ASSESSING VALIDITY, RELIABILITY, AND FAIRNESS OF QURANIC RECITATION ASSESSMENT RUBRICS INSTRUMENT IN THE MUSABAQAH TILAWATIL QURAN (MTQ): LEVERAGING MANY-FACETS RASCH MEASUREMENT (MFRM)

A Thesis

Submitted to the Master's Study Program of Education in partial fulfillment of the requirements for the degree of

Master of Arts (M.A.)

By:

Muhammad Lutfi Assidiqi

04212310008

UNIVERSITAS ISLAM INTERNASIONAL INDONESIA
DEPOK
2025

ASSESSING VALIDITY, RELIABILITY, AND FAIRNESS OF QURANIC RECITATION ASSESSMENT RUBRICS INSTRUMENT IN THE MUSABAQAH TILAWATIL QURAN (MTQ): LEVERAGING MANY-FACETS RASCH MEASUREMENT (MFRM)

A Thesis

Submitted to the Master's Study Program of Education in partial fulfillment of the requirements for the degree of

Master of Arts (M.A.)

By:

Muhammad Lutfi Assidiqi

04212310008

UNIVERSITAS ISLAM INTERNASIONAL INDONESIA
DEPOK
2025

ACKNOWLEDGMENT

All praise be to Allah, the Most Gracious, and the Most Merciful.

With His permission and grace, every page in this thesis can be written and completed.

To all souls whose kindness and presence have left an imprint on this journey,

My deepest gratitude to Dr. Lukman Nul Hakim, my academic advisor and thesis supervisor, who has trusted and allowed me to grow and do what I believe in. Through the knowledge that has been given, I move more confidently during my master's studies.

To Bambang Sumintono, Ph.D., my thesis supervisor, but also my Rasch inspirator, who gently opened the door of opportunity to Rasch measurement. A science I had never known would captivate me so deeply. A language of precision and beauty that I have come to love. This thesis is a tribute to the knowledge that you have given me. May it be the beginning of something lasting.

Special thanks were also given to my examiners, Soeharto, Ph.D., and Bahrul Hayat, Ph.D., who were willing to read and provide such invaluable suggestions for the improvement of my thesis. To all my beloved lecturers at the Faculty of Education, UIII: Dr. Destina, Assoc. Prof. Tati, Prof. Zuhdi, Assoc. Prof. Chary, Dr. Amich, Prof. Nina, Prof. Bambang Suryadi, Prof. Sahin, Dr. Suparto, Dr. Andar, Dr. Alpha, and to the dedicated faculty staff, thank you for the wisdom, sincerity, and knowledge you have shared. Your presence has helped me grow far beyond this degree. Being part of this faculty is a life decision that I will never regret; it is one of the greatest blessings I have ever received. And to my dear friends, MA Batch 3, every late night of our doing assignments, every shared meal order, every quiet walk under the night sky back to the dorm, these are not just memories, but treasures I will carry with me, always.

To my dearest family: my mom, my late father who now rests in heaven, and my beloved siblings, Mba Dini, Diyah, and Difa, thank you for believing in me, praying for me, and faithfully accompanying me in every step of my study journey.

This thesis holds your names between every line.

ABSTRACT

Muhammad Lutfi Assidiqi
04212310008
Muhammad.assidiqi@uiii.ac.id
Education
Universitas Islam Internasional Indonesia

Psychometric attribute testing of an assessment instrument is a crucial step that must be undertaken before the instrument is widely implemented. Failure to implement this stage can increase the risk of assessment discrimination, leading to unfair outcomes and a loss of credibility in the evaluation results. One important instrument that has yet to be empirically validated is the Quranic recitation assessment rubric in the Musabaqah Tilawatil Quran (MTQ) competition. Although some normative evaluations have been conducted, quantitative evidence regarding its validity, reliability, and fairness remains limited. This gap helps explain the recurrence of injustice and participant dissatisfaction with the assessment results in the MTQ competition in recent years. Moreover, the complexity of the assessment system, characterized by multiple raters and a performative evaluation, has not been fully accommodated by its current analytical approaches. Therefore, this study aims to empirically examine the quality of the Quranic recitation assessment rubric using the Many-Facet Rasch Measurement (MFRM) approach. Adopting a quantitative and non-experimental design, the study involved 50 students as the ratees and 16 judges as the raters. The assessment was conducted using the official rubric from the Tilawatil Quran Development Institute (LPTQ), which consists of four dimensions: *Tajwed, Fashahah, Lagu,* and *Suara*, comprising 19 items. To ensure assessment consistency, raters received a workshop and a guidebook before the assessment process. The data collection produced 3.760 quantitative responses and 674 qualitative responses, with 40 responses identified as invalid. The results demonstrate that the rubric possesses good construct validity, with Infit and Outfit MnSq values ranging from 0.98 to 1.27 and 0.78 to 0.98, respectively. Infit and Outfit ZSTD values ranged from -0.3 to +1.3 and -0.5 to -0.3, respectively. While point measure correlation values ranged from 0.35 to 0.56. Reliability was also found to be satisfactory, with values ranging from 0.52 to 0.88 for ratees, 0.92 to 0.99 for items, and 0.87 to 0.92 for raters. Fairness was found to be relatively high, as both ratee and rater data fit the model, with only 0.02% significant bias and 4.06% unexpected responses. However, one invalid item was identified in the *Fashahah* dimension, along with several disordered thresholds in the middle scale categories. Additionally, the reliability and separation index for the ratees in the *Fashahah* dimension were statistically low. These findings highlight the need for revision of certain item descriptors and restructuring of the rating scale categories. Further training for raters may help align their interpretations of the rubric and promote more standardized and fairness in assessments.

Keywords: *Fairness, Many-Facet Rasch Measurement, MFRM, MTQ, Musabaqah Tilawatil Quran, Quranic recitation, Reliability, Rubric, Validity.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDIX

ABBREVIATION DIRECTORY

AERA	*: American Educational Research Association*
APA	*: American Psychological Association*
BC	*: Before Christ*
CE	*: Comman Era*
CTT	*: Classical Test Theory*
d.f.	*: Degree of Freedom*
DIF	*: Differential Item Functioning*
DIHQA	*: Dubai International Holy Quran Award*
DRF	*: Differential Rater Functioning*
EFL	*: English as a Foreign Language*
GSKK	*: Geng Seni Kayu Kreatif*
IRT	*: Item Response Theory*
JQH	*: Jam'iyatul Qurro wa Al-Huffadz*
LPPN	*: Lembaga Pengembangan Pesarawi Nasional*
LPTQ	*: Lembaga Pengembangan Tilawatil Quran*
MFRM	*: Many-Facet Rasch Measurement*
MNSQ	*: Mean Square*
MoRA	*: Ministry of Religious Affairs*
MSE	*: Mean Square Error*
MTQ	*: Musabaqah Tilawatil Quran*
NCME	*: National Council on Measurement in Education*
NU	*: Nahdhatul Ulama*
PCK	*: Pedagogical Content Knowledge*
PCM	*: Partial Credit Model*
PESPARANI	*: Pesa Paduan Suara Gerejani*
PESPARAWI	*: Pesta Paduan Suara Gerejawi*
QVCI	*: Quranic Verbal Communication Index*
RMSE	*: Rasch Measurement Standard Error*
RSM	*: Rating Scale Model*
S.D.	*: Standard Deviation*
SACC	*: Scale of Aesthetics and Creativity in Chess*
Saw.	*: Shalallah 'Alaih wa al-Salam*
SciTS	*: Science of Team Science*
SKB	*: Surat Keputusan Bersama (Joint Decree)*
STQH	*: Seleksi Tilawatil Quran wa al-Hadith*
Swt.	*: Subhanah wa Ta'ala*
ZSTD	*: Z Standardized*

CHAPTER I
INTRODUCTION

Musabaqah Tilawatil Quran (MTQ) is a Quran reading competition that aims to preserve the tradition of Quran reading education in Indonesia. This competition is unique because it combines theoretical aspects of Al-Quran knowledge with aesthetic elements through the beauty of reader voices, sounds and melodies. Accordingly, the MTQ competition is governed by a multifaceted assessment framework. However, the dynamics and controversies of the competition results often cause debate among MTQ participants. Allegations of bias, concerns over transparency, and accusations of politicization of competition results have been raised repeatedly. These issues indicate problems in the evaluation and assessment process, which should be fair and objective, highlighting the need for improvements in the system. To address these challenges, this research intends to propose the use of an approach that can identify and overcome system weaknesses so that it can minimize evaluation and assessment issues in MTQ competitions. In more depth, this chapter will detail the research background, research questions, research objectives, and research significance, to provide a complete, systematic understanding.

1.1. Research Background

1.1.1. The Growing Importance of Quranic Recitation

Since the revelation of the Qur'an in 610 C.E. from Allah Swt. to the Prophet Muhammad Saw. through Jibril AS. in the Arabic language (Al-Qattan, 1995), the tradition of studying the Quran has persisted to the present day. This tradition encompasses a range of practices undertaken by numerous Muslims, including reading, memorization, comprehension, writing, and interpretation (Shihab, 2002). A distinctive practical tradition of the Quran is *Tilawah Al-Quran*, which refers to the recitation of the Quran in its original language. This practice incorporates the aesthetic dimensions of tone and voice while adhering to the established rules of reading (Indra, 2019; Nelson, 2010; Salim, 2004). The practice of *Tilawah Al-Quran*, or what is referred to in this study as 'Quranic recitation', is consistent with the prophetic guidance outlined in the Quran Surah Al-Muzammil verse 4, "... *and recite the Quran 'properly' in a measured way*," and Quran surah Al-Furqon [25] verse 32, "... *and We have revealed it in a deliberate pace*." This practice is further substantiated by a hadith from the Prophet Muhammad narrated by Abu Dawud No. 1468 and An-Nasa'i No. 1015, which states, "Adorn the Quran with your voices" (Amrullah, 2001; Abu Dawud, 1998; An-Nasa'i, 1991). The necessity of reading the Qur'an beautifully is further

substantiated by the emergence of *nagham* science, which studies various reading melodies that can be utilized in the recitation of the Qur'an (Salim, 2004). These melodies encompass various *maqam* (songs) which are called *bayyati, hijaz, nahawand, shaba, rast, jiharka*, and *sika* (Salim, 2004). According to Rasmussen (2010), an American educator and ethnomusicologist specializing in Arabic music and Islamic rituals and performances, the Qur'anic recitation has evolved into a scientific discourse, founded on religious teachings and the principles of *nagham* science, which continue to be studied and preserved (Rasmussen, 2010).

The global Islamic community has made concerted efforts to reinvigorate the scientific discourse concerning the recitation of the Qur'an. One of the notable initiatives is the organization of Quran recitation competitions. Examples of such practices, which are carried out by the global community in various countries, include the International Quran Recitation Competition (Recitation of Al-Quran Antarbangsa) in Malaysia, the Dubai International Holy Quran Award (DIHQA) in Dubai, the King Abdul Aziz International Quran Competition in Saudi Arabia, and the Musabaqah Tilawatil Quran (MTQ) in Indonesia (DIHQA Organization, 2020; Ministry of Islamic Affairs Saudi Arabia, 2021; MoRA of Indonesia, 2022). The provision of prize for the winners of these competitions also encourages the motivation of many individuals to intensify and deepen the discourse of Quran recitation (Kohn, 2021; Ormrod, 2020; Ryan & Deci, 2020). Moreover, Islamic teachings encourage Muslims to engage in a spirit of competition in the pursuit of goodness, as evidenced by the Quran in Surah Al-Baqarah [2] verse 148: "...*so compete with each other in doing good*...". Therefore, it is not surprising that many global communities are reviving the science of Quran recitation, one of which is through the organization of competitions.

1.1.2. Musabaqah Tilawatil Quran (MTQ)

In Indonesia, “Musabaqah Tilawatil Quran (MTQ)” is a Quran reading competition that was officially held and became a national celebration after the issuance of “The Joint Decree (SKB) of the Minister of Religion and the Minister of Home Affairs of the Republic of Indonesia Number 19 of 1977 / Number 151 of 1977 concerning the establishment of the Tilawatil Quran Development Institute (LPTQ)”. LPTQ is an informal educational institution under the responsibility of the Ministry of Religion. The institution is responsible for organising MTQ and providing Quranic education. The education begins with the study of reading, memorizing, understanding, interpreting, translating, and then

appreciation and practice for the society ("Decree of the Minister of Religion Affairs of the Republic of Indonesia Number 240 of 1989 Article 3"). Referring to the book published by the MTQ Foundation in 1989 written by Gani, this event was carried out by Jam'iyatul Qurro wa Al-huffadz (JQH) under the auspices of Nahdhatul Ulama (NU) as a form of gratitude for the independence of the Indonesian nation in 1945 (Baharuddin et al., 2022; Kemenag, 2022; Salim, 2004). This initiative was traced to have been carried out in North Sumatra on February 12, 1946, although on a limited scale (Gani, 1989).

Following the implementation of the policy, LPTQ has maintained its momentum, particularly in the organisation of MTQ competitions. According to the records of the Ministry of Religious Affairs, the MTQ has been held until the 30 times in 2024 (Kemenag, 2024). A similar competition, known as the Selection of Tilawatil Quran and Hadith (STQH), was held for the 27 times in 2023 (Kemenag, 2023). As a national event, the MTQ is routinely carried out every year, gradually from the sub-district to national levels (MTQ Guidebook, 2023). Consequently, it is unsurprising that MTQ activities are carried out in almost all regions in Indonesia. The management of LPTQ, which is now spread across 38 provinces, 416 districts, 98 cities, and 7.285 sub-districts, is under the jurisdiction of the Ministry of Religion and local governments (Decree of the Minister of Home Affairs Number 300.2.2-2138, 2025). This underscores the notion that the execution of the MTQ competition serves as a national celebration, exerting a profound and extensive influence.

The objective of the MTQ is not confined to a competitive event; it is, in essence, an inclusive educational forum designed to promote self-development, particularly in the science of the Quran. Thus, a variety of practice models have been implemented, including numerous competitions encompassing diverse disciplines such as Quran recitation (*Tilawah, Tartil, Murotal Qira'at Sab'ah*), memorization (*Hifdzil Quran, Hifdzil Hadith*), comprehension (*Fahmil Quran, Tafsir Al-Quran*), writing (calligraphy and *khat*), and scientific studies of the Quran (Scientific Writings competition) (MTQ Guidebook, 2023). These competitions are open to all groups, including children to teenagers, men and women, and people with special needs (MTQ Guidebook, 2023). Furthermore, the MTQ has sought to integrate learning principles, competencies, and spiritual aspects (MTQ Guidebook, 2023). Spiritual here refers to efforts to instil the values contained in the Qur'an through various categories of competitions such as the desire to continue to study the Qur'an (Qs. *Thaha* [20]: 114), compete in goodness (Qs. *Al-Baqarah* [2]: 148), to uphold the practice of justice and sportsmanship, especially in the implementation of competitions (Qs. *Al-Maidah* [5]: 8; *Al-Nahl* [16]: 6; *Al-Nisa* [4]: 58, 135; *Al-Sad* [38]: 26). These

practices can be regarded as forms of inclusive education, seeking to be embedded in the society through the cultivation of Quranic-based character education in MTQ activities.

1.1.3. Assessment Issues in the MTQ Competition

The MTQ competition, organised by LPTQ, is an educational activity intended to inculcate the values of the Qur'an (Ministry of Religion & Ministry of Home Affairs, 1977). All practices must align with the Qur'an, including its educational values, from preparation to the culmination of activities. As a competition, one of the most important values to be pursued is fairness and honesty (Sutter & Kocker, 2004). The principles of measurement and assessment in education also affirm these values, ensuring that discrimination is avoided (Wilson, 2005). The assessment system is recognised as the most vulnerable area (Sutter & Kocher, 2004; Plessner & Haar, 2006; Dawson & Dodson, 2010). Efforts must be made to ensure that the systems and practices of the MTQ competition align with the Quranic values.

In the context of the MTQ competition, the combination of Quranic recitation's technical science and its aesthetic or artistic vocalization aspects has made the assessment process complex. This complexity is proven by the MTQ manual book that each participant will be assessed by 12 judges, divided into four different groups. Each group which consists of three judges is responsible for assessing one of the four dimensions. The four dimensions - *Tajwed, Fashahah, Lagu,* and *Suara* - each have four to five items to be assessed with different criteria of standards. Each dimension has a minimum to maximum score. The score is based on the errors or imperfections in the Quranic recitation. Guidelines and rules regarding the assessment threshold between judges have also been set in detail, including technical rules for determining the winner. All explanations of this system are included in the "Al-Quran and Al-Hadith Musabaqah Guidebook," published in 2023 by the Directorate of Islamic Religious Information and the Directorate General of Islamic Community Guidance, Ministry of Religious Affairs.

The assessment system has been made quite detailed and strict. However, issues related to the assessment that impact the competition winner's decision still often arise in this competition. According to newspaper articles published by "IDN News Lampung" (2023), "Infokini News" (2024), and "Totabuan News" (2024), the credibility of judges is always a source of fraudulent issues in the assessment due to internal and external factors. Internal, can be understood as a misunderstanding of the assessment rubric, a lack of knowledge about the dimensions being assessed, or a lack of training received. While

external factors can take the form of encouragement from other parties, such as politicization by the host organizer, relationships between the judges and participants, or efforts to ensure the success of certain parties. These factors can reasonably impact the competition results, which indirectly do not reflect the quality decision that should be.

The organizers have begun to address the issue of unfair assessments. Since 2016, the LPTQ, through a report provided by the Ministry of Religion, has encouraged the use of digitalization in the national MTQ competition level in West Nusa Tenggara (Ali, 2016). However, digitalization has only been implemented for participant registration to ensure that there is no participant who does not meet the requirements registered. Regarding assessment, the effort of live scores after participants' performances is only optimally carried out at the national level and in several regions with adequate technological capacity and funding.

In 2014, a study conducted by Bahruddin and Kumaidi (2014) also offered a critical evaluation of the MTQ assessment and found that the problem stemmed from the assessment system, which was still problematic, including the rubric. There were at least four reasons put forward. First, the assessment instrument used was less fair; second, the assessment rubric was less proportional; third, the assessment model tended to be inconsistent; and fourth, there were limitations in the independence of the assessors (Bahruddin & Kumaidi, 2014). To address this issue, a research and development approach was used, in which the assessment system and rubric were reconstructed by considering the four weaknesses identified in the original system. The findings obtained from this reconstruction showed a high reliability value ranging between 0.96 and 0.97. However, the analysis has not yet measured and quantitatively assessed the rubric that was still in use at that time. In the context of rubric assessment, a quantitative approach is important to ensure that the criticisms built are more valid and have strong empirical evidence. Therefore, although the LPTQ has provided a digitalization in the system of the competition, and Bahruddin and Kumaidi's (2014) research has provided valuable efforts, these critics show conjecture and clarity on the reasons why problems related to assessment persist in the following years, regardless of the various underlying factors.

The emergence of the issues has had a significant impact on the appearance of judgment biases that contravene the fundamental principles of measurement and assessment (Engelhard, 2013), including the values enshrined in the Quran. The assessment bias exhibited by judges in the field of assessment is theorized to be influenced by their

respective backgrounds, thereby giving rise to phenomena of rater effects, such as severity or leniency in scoring (McNamara, 1996). The presence of such bias can be attributed to the presence of an invalid and unreliable assessment system or rubric (Wilson, 2023). Controversies and protests are inevitable if the integrity of the judges as assessors and/or the quality of the assessment rubric are problematic, because they can provide results that do not reflect the performance of the participants. In addition, because of the incorporated assessment, the credibility of the competition may be affected, which can also negatively impact participants' motivation to keep learning and competing (Carmona-Halty et al., 2021; Engels et al., 2021; Tierney et al., 2013). Therefore, given the general implications of this competition, it is necessary to conduct a study to evaluate the assessment system utilized in the MTQ.

1.1.4. A Tool for Evaluating MTQ Assessment

In the field of assessment, continuously over time, various approaches to produce valid and fair assessments have developed (Bond & Fox, 2015; Boone et al., 2016; Sumintono & Widhiarso, 2013). In 1904, Charles Spearman developed the Classical Test Theory (CTT), Frederic M. Lor and Allan Birnbaum developed the Item Response Theory (IRT) in the 1940s, until the emergence of the Rasch Model in the 1960s, developed by George Rasch (Wang & Osterlind, 2013; Bond & Fox, 2015). The Rasch Model is a probability-based measurement with the assumption of unidimensionality designed to measure the relationship between participant ability and the item level difficulty (Rasch, 1960). No matter who is being tested or what items are used, its sample-independent nature enables consistent measurement (Bond & Fox, 2015). In contrast to the Classical Test Theory (CTT), which is limited to the ability to measure certain samples and has not been able to identify latent information (Wang & Osterlind, 2013; Crocker & Algina, 2006). Even though the Rasch model, which is part of IRT, can handle more complex situations than CTT, this model is only suitable for unidimensional data and is unable to handle complex measurement, such as a multidimensional situation (Bond & Fox, 2015; Traub, 1994).

In 1989, John Michael Linacre introduced a new model which is an extension of the Rasch model. This model is called Many-Facet Rasch Measurement (MFRM) to overcome the complexity of data that has several aspects of measurement dimensions (Linacre, 1989). The involvement of several aspects of measurement, or what is known as facets, in the MFRM approach, can be analysed separately to identify their contribution to the measurement results. The analysis can be used to reveal assessor bias (Severity or leniency,

inter-rater consistency, central tendency, halo effect) (Engelhard, 2013), rubric quality (valid and reliable) (Wang & Osterlind, 2013; Myford & Wolfe, 2003), and identification of the distribution of assessment items, the panel of judges, and participant abilities (Eckes, 2011).

To respond to the controversy occurred in the results of the MTQ assessment, the MFRM analysis model can be a solution to address this phenomenon. The analysis offered can measure all facets independently, so that it allows for a more detailed description of the MTQ issue from various perspectives (Linacre, 1989; Engelhard, 2013). These results can be obtained from the judges who assessed the participants using the Quranic recitation rubric assessment. Therefore, this research intends to test the quality by checking the validity, reliability, and capacity of the rubric in providing fair assessment by utilizing the MFRM analysis. The goal is to provide empirical evidence, either rejection or support for the issues that occur in the MTQ assessment, with a more measurable, robust, and comprehensive approach.

1.2. Research Questions

MFRM is a psychometric analysis model that evaluates assessments involving multiple raters (Linacre, 1989). Through its analytical ability to separate the influence of judges, participants, and assessment rubrics, MFRM is expected to reveal the quality of MTQ assessments (Linacre, 2012; Eckes, 2015). The analysis of various aspects, including the validity and reliability of the assessment rubric, the identification of distribution of participants' abilities and the difficulty level of items, as well as the impact of that, can provide valuable latent information that can be used to evaluate the rubric (Bond & Fox, 2015; Engelhard, 2013; Linacre, 2018; 2023). In line with this rationale, the following four research questions are posed:

1. To what extent does the Quranic recitation assessment rubric demonstrate good/excellent psychometric validity based on the results of MFRM analysis?
2. How reliable is the rubric in ensuring consistent scoring across different judges and participants, as evidenced by MFRM analysis findings?
3. To what extent does the assessment ensure fairness for all participants, considering judge severity, item difficulty, and unexpected scoring patterns revealed through MFRM analysis?

1.3. Research Objectives

In reference to the three questions posed, this study has three research objectives:

1. To assess the psychometric validity of the Quranic recitation assessment rubric used in MTQ competitions through MFRM analysis.
2. To evaluate the reliability of the rubric in producing consistent scores across different judges and participants, using statistical indicators from MFRM analysis.
3. To examine the fairness of the assessment process by analysing judge severity, item difficulty, and unexpected scoring patterns as revealed by MFRM analysis.

1.4. Research Significances

This research is projected to have significance in both theoretical and practical aspects. Theoretical refers to the capacity of research to contribute to scientific theory. Meanwhile, practical contributions pertain to the relevance or application of research results in real-world contexts. Here are the details:

1. Theoretically

This research can contribute to the development of theories in the field of performance-based educational assessment, especially in the context of Quranic recitation. The employment of MFRM within the assessment system of religion-based competitions has the potential to offer a new perspective on literature studies. Furthermore, these findings can provide empirical evidence and serve as a scientific basis for evaluating, improving, and/or developing a more valid and reliable Quranic recitation assessment rubric.

2. Practically

From a practical standpoint, the results of this study have the capacity to impact numerous individuals and institutions. First, the findings on the quality of the assessment rubric in terms of validity, reliability, and fairness can provide empirical evidence and constructive recommendations to improve the Quranic recitation assessment rubric for LPTQ across 38 provinces, 416 districts, 98 cities, and 7.285 sub-districts in Indonesia. The evaluation of the assessment practices carried out can possibly have implications for many groups, including participants, judges, and pedagogues. Second, this research can provide a basis understanding on the quality of the assessment rubric that has been used in the MTQ competition. Finally, educational institutions that offer Quranic recitation courses can adopt the evaluation method, and the recommendations offered to produce a good quality assessment, particularly for learning and teaching purposes.

CHAPTER II
THEORETICAL FOUNDATION

In this chapter, the theoretical foundation as the guidance of this research is provided. As outlined in Chapter I, the objective of this study is to assess the Quranic recitation rubrics in the MTQ competition by employing the MFRM analysis model. However, the investigation of the academic discourse or what can be called as the literature review, the formulation of the theoretical framework, and the conceptual framework, has not been thoroughly explained. A literature review is a process that can facilitate the establishment of context, the strength of arguments, the identification of research gaps, and the formulation of methodologies (Booth et al., 2016; Hart, 2018). Theoretical frameworks can provide a foundation while facilitating the interpretation of findings in the context of existing theories (Anfara & Mertz, 2006). The conceptual framework can be visualized as a representation of the relationship between variables, in this research between multiple facets, thereby serving as a guide in constructing the structure of analytical discussions (Ravitch & Riggan, 2017; Engelhard, 2013; Maxwell, 2013). In this section, these three components will be systematically described to explain and affirm the position of this research.

2.1. Literature Review

2.1.1. The Significance of Quranic Recitation Competition

Referring to many literature studies, Quranic recitation competitions have been widely revealed to have a massive positive impact on the wider Muslim community. Education is the area that benefits the most, especially in encouraging the productivity of the dissemination of Quranic sciences. Hussaini (2020), who conducted a study in Nigeria, found that Quranic recitation competitions have increased the discourse of Quranic studies, especially in Bauchi Metropolis in Bauch Local Government area of Bauchi State (Hussaini, 2020). There have been many changes in terms of curriculum, syllabus, and even teaching methodology in various educational institutions (Hussaini, 2020). This treatment encourages students to further study Quranic science, such as *tajwed* and *qira'at*. According to Hussaini (2015), Quranic recitation has become a new trend and dominates the Islamic education curriculum in many regions in Nigeria. Yahaya et al. (2024) confirmed that in Northern Nigeria, the implementation of Quranic recitation competitions has encouraged students' enthusiasm for learning, especially in learning Arabic to have a better understanding of the Quran and its sciences (Yahaya et al., 2024). Furthermore, Quranic

recitation as a performance-based has also been accommodated by various educational institutions with special study offers. This can be confirmed on the websites of some universities, such as Al-Azhar University in Egypt, Umm Al-Qura University in Saudi Arabia, the International Islamic University of Malaysia in Malaysia, and the Institute of Quranic Sciences in Indonesia who offered the Quranic science like *tafsir* or Quranic exegesis, recitation with *tajwed,* memorization, and the study of its sciences including *nagham* science.

Several further studies have found that the influence of Quranic recitation competitions has also expanded due to the impact and contribution made by the winners. Maria Ulfah is the first woman from Indonesia to gain widespread recognition in the world of Quranic recitation at the international level in 1980 (Rasmussen, 2010). According to Rasmussen (2010), Maria Ulfah's influence as the winner of the international Quranic recitation competition in Malaysia has created increasing attention to Quranic recitation education both formally and informally. Not only that, The Royal Islamic Strategic Studies Center (2024) also awarded Maria Ulfah for more than 10 consecutive years as the 500 most influential Muslims in the world (The Royal Islamic Strategic Studies Center, 2024). The influence given after winning the MTQ competition has succeeded in encouraging society, especially women's involvement in studying Quranic recitation (Rasmussen, 2010). In other research, Safie et al. (2021) also studied a winner of an international Quranic recitation competition named Faridh binti Mat Saman from Malaysia. The research said that Faridh's influence has also motivated many educational institutions in Malaysia to implement the study of Quranic *Tarannum* (recitation with beauty) (Safie et al., 2021).

A study conducted in Indonesia found that the MTQ competition has had a broad impact on society and the surrounding environment. Hasan's (2019) research, which conducted a historical study while exploring the opinions of scholars, found that MTQ is a positive activity that has educational value, competitive character education. Armiadi et al. (2023) also explained that the MTQ competition can improve the religious practices of participants as well as become a tool for the preaching of Islam in Indonesia. However, Harahap (2023) and Setiawan et al. (2024) revealed that the MTQ competition has weaknesses that could cause achievement motivation challenges for many participants. Harahap (2023) explained that some parents of participants tend to pressure their children to win the competition just for happiness or to get the prize, which could be not only in the MTQ context. Responding to this, Setiawan et al. (2024) attempted to promote mental health and increase participant motivation by providing treatment through workshop

activities. The results revealed that these activities were successful and could be implemented because they had been proven to significantly increase the motivation and self-confidence of 75% of participants (Setiawan et al., 2024).

Even more impressive, the MTQ competition also has an educational tolerance influence in society. Research conducted by Baharuddin et al. (2022) in the Saumlaki area, Maluku Province, in the Eastern part of Indonesia, with a Muslim population of only 4% (now 18.03%) encouraged the involvement of non-Muslims as the committee for the implementation of this competition. Non-Muslims also helped and practiced tolerance by adjusting costumes and allowing the use of the Catholic Center Hall as one of the activity venues (Baharuddin et al., 2022). Muslim communities who use non-Muslim religious places also do the same thing by not removing a Cross and a painting of the Virgin Mary (Baharuddin et al., 2022).

To this point, the practice of society's tolerance cannot be separated from the government's contribution in providing space for growth and preaching for every recognized religion in Indonesia. Like Muslim, which has MTQ, other religious adherents also have similar performance-based practice competitions. Christianity has PESPARAWI or the Church Choir Festival which is held routinely at the national level every three years under the LPPN or the National PESPARAWI Development Institute based on "The Regulation of the Minister of Religious Affairs of the Republic of Indonesia Number 19 of 2005" (Ministry of Religious Affairs, 2005). Hinduism has *Utsawa Dharmagita* or a festival of Veda scripture reading competitions under the Dharmagita Development Institute based on "The Regulation of the Minister of Religious Affairs of the Republic of Indonesia Number 28 of 2016" (Ministry of Religious Affairs, 2016). Other religions such as Catholic Christianity have PESPARANI or the Catholic Church Choir Festival and Buddhism has *Mahanitiloka Dhamma* (Ministry of Religious Affairs, 2018; 2022). These various types of competitions, which also involves the practice of reading and understanding the holy book through vocals and music, represent methods of religious education that indirectly promote tolerance among its adherents. This confirms the influence of tolerance values in the MTQ competition because it accommodates the context of a diverse Indonesian society. Therefore, from all the existing literature shows that implementing the Quranic recitation competition has many positive impacts, specifically for the global community to promote, educate, practice, develop, and encourage the preservation of the Quranic science.

2.1.2. Quranic Recitation Assessment in the MTQ Competition

Referring to the "Al-Quran and Al-Hadith Musabaqah Guidebook" published in 2023 by the Directorate of Islamic Religious Information and the Directorate General of Islamic Community Guidance, Ministry of Religion, the MTQ competition has a detailed and complex assessment system. All types competed in this competition have different systems, especially regarding the dimensions and items assessed in the assessment rubric. Quranic recitation as one of the types of the competition, has four dimensions that are measured to determine the winner. These dimensions consist of *Tajwed, Fashahah, Lagu,* and *Suara*.

The first dimension is *Tajwed.* This dimension is used to evaluates the accuracy of Quranic recitation of participant. Al-Jazari (2017) explains in "*Al-Muqaddimah Al-Jazariyyah*" that *Tajwed* refers to the rules and techniques used to read the Qur'an correctly (Al-Jazari, 2017). There are several important components in the science of *tajwed*. These components are used as assessment items in the *Tajwed* dimension of the MTQ competition. According to the guidebook of MTQ competition (2023), these components consist of *makharij al-hurf, shifah al-hurf, ahkam al-hurf,* and *ahkam al-maad wa al-qshr*. *Makharij al-hurf* is an assessment item that tests the accuracy with which *al-hurf al-hijaiyah* are read. For example, the letter of *jim* must be pronounced with the tip of the tongue touching the palate. *Shifah al-hurf* assesses the fulfilment of the characteristics of reading *al-hurf al-hijaiyah* pronouncation. For example, the *hurf ta'* must be pronounced with a hiss. *Ahkam al-hurf* assesses the implementation of reading laws, such as *idzhar, ikhfa, iqlab*, and *ghunnah*. *Ahkam al-mad wa al-qasr* assesses the accuracy of reading the Quran, especially words that require to be read with specific lengths.

The second dimension is *Fashahah*. This dimension evaluates participant's fluency in reciting the Qur'an. Al-Nassir (1985) explains in his dissertation "*Sibawayh the Phonologist: A Critical Study of the Phonetic and Phonological Theory of Sibawayh as presented in His Treatise Al-Kitab*" that *Fashahah* is used as a guideline of accuracy in the articulation of letters (*makhraj*), words, and phrases of the Quran. In the MTQ competition, the *fashahah* is assessed in five assessment categories: *ahkam al-waqf wa al-ibtida, mura'ah al-hurf wa al-harakah, muro'ah al-kalimah, muro'ah al-ayah, tamam al-waqt* (MTQ Guidebook, 2023). According to the guidebook of MTQ (2023), *ahkam al-waqf wa al-ibtida'* assess the place where a recitation of the Quran stops and starts. This is important because mistakes in starting or ending the recitation, especially in the middle of a verse, can potentially change the meaning of the verse (Al-Nassir, 1985). *Muro'ah al-hurf aw al-harakah* assesses the accuracy of the reading in terms of letters and harakat. For example,

the letter *kha* must be read as *kha*, the *harakah al-fathah* as *fathah. Mura'ah al-kalimah* monitors the possibility of participants missing a word. *Muroah al-ayah* monitors the same possibility, but in terms of verses. For example, a participant who should read verse 160 suddenly missed it. *Tamam al-waqt* assesses participants' time management when reciting the Quran.

In the dimension of *Lagu* or song in the English term, the participants will be assessed on the aesthetic quality of their voices when reciting the Qur'an. This includes the use of certain maqams (songs) that align with the Quranic tradition (Salim, 2004). The *maqamat of Al-'Arabiyyah*, which includes *bayati, nahawand, shaba, rast, sika, jiharka*, and *hijaz,* are considered the only acceptable *maqamat* (Salim, 2004; Rasmussen, 2010). The suitability of utilizing *maqamat* when reciting the Quran is also considered (Rasmussen, 2010). The dimensions of *Lagu* in the MTQ competition are defined by the capacity of participants to utilize the introductory and concluding songs, the number and/or composition of songs (specifically in the final round), transitions, complete form, and tempo of the song, rhythm, style and appreciation, variation (MTQ Guidebook, 2023).

The dimension of *Suara* or voice in the English term is defined as the vocal quality exhibited by participants during the recitation of the Quran, encompassing both technical and aesthetic elements (Rasmussen, 2010). In Rasmussen (2010) research on the Quranic recitation tradition, she elucidates that sound quality is frequently associated with factors such as clarity, resonance, and respiratory control (Rasmussen, 2010). In the assessment of the MTQ competition, the dimension of *Suara* consists of five items: vocals and voice integrity, voice clarity, smoothness/softness, loudness, breathing regulation (MTQ Guidebook, 2023).

In the assessment process, which encompasses all dimensions and incorporates criteria items, is complex. Assessors are tasked with evaluating the technical and aesthetic capabilities of each participant concurrently during each performance, which has a duration of approximately 6-12 minutes, contingent on the category of participants (MTQ Guidebook, 2023). The disparity in the score threshold for certain items, and the implementation of a maximum score rule in each dimension, serves to further rigidify the assessment process. The aesthetic aspects, which are susceptible to subjectivity and depend on the characteristics of the assessor and the environment, is also a critical issue that requires careful consideration (Scherbakova et al., 2025). Another salient issue pertains to the variation in the verses evaluated across participants, a factor that demands careful consideration (Bahruddin, 2014). Recognizing the challenges, LPTQ, as the organizer of

the competition, has stipulated that each dimension must be evaluated by a minimum of three judges. Consequently, the assessment of each participant necessitates the involvement of 12 judges, who collectively assess the participant's performance based on four dimensions assessed.

In the context of assessment, multi-rater can be understood as a measurement method which offers numerous advantages, particularly in performance-based assessments like in the MTQ competitions. According to Linacre (1994), this method can increase the reliability of the assessment because the final score is the result of aggregation from various points of view. Research in 2015 conducted by Gynnild on assessing vocal performances proved that conditions. Bond and Fox (2015) also highlighted that multi-rater method can mitigate the impact of individual judgment bias. Engelhard (2002) also emphasized the importance of multi-rater because it can encourage fair assessment, especially when overcoming differences in perception or standards between raters. Wright and Stone (1979) also mentioned that reducing data distortion due to individual perception can enhance the validity of the assessment. Furthermore, differences in perspective can provide detailed insights regarding the strengths and weaknesses of assessments (Sadler, 1989).

Despite the benefits that can be gained from the involvement of multi-raters, the context of MTQ competitions that assess technical, artistic, and aesthetic elements needs to be handled properly. Previous research on similar competitions has emphasized the difficulty of involving many judges in vocal, aesthetic, artistic, or creativity-based competitions. A study by Mitchell (2014) examining the perception, evaluation, and communication of singing voices found that subjectivity in vocals and aesthetics is unavoidable. Similarly, Scherbakova et al. (2025), who examined creativity and aesthetics in chess, showed that subjective assessment is an innate trait due to differences in the assessor's background and environment. Research on vocal performances conducted by Gynnild (2016) further emphasized the need for a good rubric as well as a criterion-based assessment framework to provide clear standardization of its assessment. Alvarez-Diaz et al. (2021) who studied performance in a musical contest explained that discrepancies between expert judges are likely to occur so that the use of a clear rubric is recommended to minimize discrepancies and improve the quality of assessment. Ogari (2020) emphasized in his dissertation, through the standardization of a clear assessment framework (rubric) can create a fairer, more comprehensive, and thorough assessment.

Despite its meticulous design to ensure comprehensive, objective, and unbiased evaluations, the MTQ assessment system has numerous inherent weaknesses, as identified

by numerous sources. Azwar's (2018) research proposing the reconstruction of MTQ implementation was motivated by his findings on the frequent occurrence of empirical problems in the MTQ competition. These problems included manipulation of participant identities, cheating between officers, and indications of non-transparent and non-objective assessments by the judges because political figures wanted to win for their host region (Azwar, 2018). These issues are still emerging and continue to be reported in several regions until 2024 ("IDN News," 2023; "Infokini News," 2024; "Totabuan News," 2024; "Metronews.co," 2024). According to Azwar (2018), this condition must be addressed immediately through a comprehensive reconstruction to align it with the values outlined in the Al-Quran-*Rahmatan li al-'alamin*.

In 2016, MTQ organizers have made efforts to provide transparent assessments by using e-MTQ technology. Ali (2016) who reported this application has conducted observations, documentation, and direct interviews during the implementation of the national MTQ in West Nusa Tenggara. Ali revealed that the existence of technology has impacted transparency and accountability in assessments by reducing the number of participants who do not comply with national regulations. Not only that, by providing live score reports imediately after participants complete their performence, these reports are considered usefull in overcoming the potential for unfair assessment. However, judges training and technological infrastructure development are still concerns that need improvement, particularly when dealing with traditional judges and regions lacking adequate technological capacity and financing for implementation Therefore, it is not surprising that the subsequent studies (still) have found problems in the MTQ competition.

In 2014, Bahruddin's (2014) research conducted a critical analysis of the MTQ assessment system. An investigation was conducted into the weaknesses of the assessment system which were then addressed by revising the assessment rubric by referring to the Borg and Gall (1983) procedure. The objective of this research was to identify the most suitable assessment rubric formula that is more aligned with the principles of measurement and assessment, namely validity, reliability, and fairness. The research utilized a research and development design to recreate the rubric, resulting in a relatively high reliability efficiency of 0.96, thereby demonstrating the reliability of the developed model (Bahruddin & Kumaidi, 2014). Unfortunately, these efforts have not yet been incorporated into the implementation of MTQ.

The research conducted by Bahruddin and Kumaidi (2014) is commendable as it proposes a more reliable scoring system. Their argument is reasonable, which is based on

criticism of the assessment rubric found in literature studies. However, it is important to note that their research had not yet tested the assessment rubric used in the MTQ competition, especially in giving empirical support of the rubrics' quality. Whereas, by providing empirical findings, robust empirical evidence can be offered and the issues in the MTQ assessment can be solved by accurate identification (Creswell & Cresswell, 2018). For this reason, before developing a better Quranic recitation assessment rubric, the current assessment rubric used must be tested first.

2.1.3. Key Psychometric Concepts in Assessment

To address the needs of the Quranic recitation assessment rubric testing, a study of the fundamental aspects that underpin ideal measurement and assessment must be explored. Referring to Engelhard's (2013), Sumintono and Widhiarso (2015), and Wind (2018), there are three fundamental areas in evaluating measurement and assessment. These three fundamental areas are validity, reliability, and fairness, as standard for instrument testing formulated by American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). These principles were later codified in the Standards for Educational and Psychological Testing, which for the new version was published in 2014.

The term 'validity' refers to the extent to which a measuring instrument can measure what it is intended to measure (The Standards, 2014). This includes theoretical and empirical evidence that supports the interpretation and use of measurement results (The Standards, 2014). In the context of MTQ, rubrics that have four dimensions of assessment must be able to accurately measure the ability of Quranic recitation. If the rubric emphasizes more aesthetic aspects without paying attention to technical aspects such as *Tajwed* or *Fashahah*, then its validity is questioned.

Reliability is defined as the consistency of measurement results when repeated under similar conditions (The Standards, 2014). It encompasses stability, internal consistency, and the similarity between raters (Engelhard, 2013). In the context of MTQ competition, if two or more judges assess participants differently for the same *tajwed* ability, then the reliability of the assessment is questionable. Similarly, if the assessment rubric used for the next competition yields different results for similar participants abilities, then the reliability of the rubric is called into question. Statistically reliability refers to volume of data and its variations, when more judges assess many participants, it will be resulted more volume data which can inform its reliability better.

In assessment, fairness signifies the absence of irrelevant biases, thus ensuring equal opportunities for all participants to demonstrate their abilities (The Standards, 2014). An assessment rubric should be designed to minimize bias caused by demographic, linguistic, cultural, or other background-related factors (Engelhard, 2013). In the context of MTQ, the background or recitation style of participants should not have a negative impact on the assessment process. However, if the panel of judges tends to favour a particular *maqamat* due to personal bias rather than adhering to the assessment guidelines, which are not utilized by participants from a specific region, then the assessment process must be re-examined.

As Wilson (2023) emphasizes, these three aspects are important components for demonstrating the quality of an assessment rubric. Research on the need to create good assessment rubrics, in terms of theory and application, has been carried out and continues to develop (see, for example, Smith et al., 2021; Jones, 2022; Brown, 2023). The late 20th century marked the popularization of rubrics in education as a response to the need for standardized, measurable assessments (Sadler, 1989). The provision of explicit assessment criteria to support formative learning is particularly important for providing feedback to students (Sadler, 1989). Therefore, assessment rubrics must bridge the gap between student performance and expected standards.

In the early 2000s, researchers began to highlight the importance of the validity and reliability of rubrics as educational assessment instruments. Moskal and Leydens (2000) discussed the fundamental principles that a rubric must valid and reliable. The valid and reliable include clarity of criteria, consistency of use, and its ability to differentiate between different levels in equal-interval distance (Moskal & Leydens, 2000). In the context of educational setting, research conducted by Andrade (2005) also showed that rubrics are not only a guide for assessors but can be an effective evaluating tool for students since the feedback can be formulated accurately. According to his findings, a well-designed rubric can provide students with an understanding of learning objectives which then impacts their performance results (Andrade, 2005). Citing Marzona's (2006) terminology that examines the history of assessment in American education, producing 'meaningful feedback' is the goal.

In the following decade, attention to the quality of rubric assessment shifted to the use of statistical methods to validate assessment rubrics. For instance, research conducted by Ginkel et al. (2017) who designing a rubric instrument in the oral presentation performance by testing its validity through an expert group. Brookhart (2013) explains that,

along with content validation, it is important to assess construct validation to ensure that the rubric accurately measures the intended aspect. Research conducted by Wesolowski (2018) dan Wesolowski and Wind (2020) emphasized that the shift in understanding validity from content validity-criterion validity-to construct validity is due to the existence of unobservable latent factors that impact the validity of an instrument. One approach that has gained considerable traction is the Rasch model, introduced by George Rasch and popularized by his colleague, Benjamin Wright (Bond & Fox, 2015; Sumintono & Widhiarso, 2015). The Rasch model in the beginning is an analysis tool used to process dichotomous data by combining algorithms that express probabilistically expected outcomes from respondent ability and item difficulty (Bond & Fox, 2015). The Rasch model also continues to discuss polytomous data types developed by David Andrich from Australia (Andrich, 1988). Not only that, The Partial Credit Model also (PCM) emerged as an analysis model for multiple rating data developed by Geoff Master from Australia for dichotomous and polytomous data simultaneously that have different ratings on some items (Engelhard, 2013).

Another model developed for analysing multi-rater data is the model developed by Mike Linacre for multiple rater data analysis, called as the Many-Facet Rasch Measurement (MFRM) (Linacre, 1989). Depending on the type of data analysed, this model can examine each aspect of the assessment individually while still reporting the psychometric attribute of the data. Thus, the possibilities of bias or inconsistency that impact to the quality of the assessment can be performed (Linacre, 1989). This is a significant point, considering Popham's (2019) statement that educational assessments should be unbiased, regardless their identity, social background, culture, and language, but not limited. A thorough review of the relevant literature reveals that validity, reliability, and fairness are the main psychometric assessments to consider when testing an assessment rubric.

2.1.4. Applications of Many-Facet Rasch Measurement (MFRM)

A review of the existing literature on the development of analytical tools for evaluating key psychometric measurement revealed an approach that can evaluate Quranic recitation assessment rubric in the MTQ competition. The Many-Facet Rasch Measurement (MFRM), which as previously explained, can accommodate Quranic recitation assessment formats involving many judges/raters that assess multiple participants (Linacre, 1989). Additionally, given that the weighting criteria of certain items in Quranic recitation rubric vary from those of others, data criteria that are regarded as dichotomous, polytomous, or a combination of both, MFRM can still be used to this condition. However, it is important to

explore how earlier researchers employed MFRM in their studies for assessing the quality of assessment, particularly in performance-based assessments like Quranic recitation.

The use of Many-Facet Rasch Measurement (MFRM) as a tool for evaluating the quality of assessment (validity, reliability, fairness-i.e., inter-rater bias) has grown rapidly in recent decades. Maryati et al. (2019) conducted a study using MFRM to assess prospective teachers' performance in terms of Pedagogical Content Knowledge (PCK). They developed a rubric which was then evaluated by MFRM through multiple raters. The study revealed that MFRM can produce the results of the validity, reliability, separation, and unidmensionality (Maryati et al., 2019). Similar research was also conducted by Yudha (2020) who provide psychometric findings by testing the validity and reliability of a performance assessment geometry rubric using MFRM. However, a more in-depth analysis of the infit and outfit mean square (MNSQ) is yet to be conducted, whereas it is needed to strengthen and enrich the findings (Engelhard, 2013). As previous studies conducted by Aryadoust (2015) and Wind et al. (2016), examining the performance of oral presentations by first-year university students and music performance assessment, respectively, found that MFRM can reveal valuable information through data fit analysis. That indicator report can be used to identify relevant or irrelevant variables in the assessment, improving its accuracy and quality (Aryadoust, 2015; Wind et al., 2016).

On the other hand, Zabidi et al. (2021) took an interesting study by conducting intercoder reliability to evaluate the content validity of codes in qualitative study transcripts using MFRM. They developed, tested, and enhanced the code scheme using the results of the comparability of each rater that could be accommodated by MFRM. Analysis at the rater level has been found to provide meaningful information, particularly in explaining rater's systematic patterns which may lead to the bias or inconsistency practices (Wu & Tan, 2016). Therefore, it is useful to compare the rater performance, items, ratees, and even rating scales as it can provide more accurate information regarding its influence on the test quality (Connelly et al., 2016; Wind et al., 2016; Zabidi et al., 2021).

A study by Wang et al. (2022) on interdisciplinary team collaboration, proves that MFRM provides valid and objective measurements in the field of the Science of Team Science (SciTS). The model calibration ability allows comparison between facets, particularly in their research who considering demographic backgrounds (Wang et al., 2022). Another practice in showing the advantages of using MFRM was conducted by Afifi et al. (2023) who developing the Quranic Verbal Communication Index (QVCI) instrument by utilizing its ability in analysing the expert validation, while still used the Rasch rating

scale model (RSM) for test validity and reliability. The latest research in 2025 by Scherbakova et al. (2025) also used MFRM in a chess competition to develop and test the validity of the Scale of Aesthetics and Creativity in Chess (SACC).

In recent years, research attention has also been given to the MFRM's capacity in identifying assessment fairness. Kudiya et al. (2018) applied MFRM to a study of Batik artisan's judgments and found that some items created unexpected responses from raters which potentially lead to unfairness. Aslanoğlu and Şata (2021) also realized that giving someone's performance a score that is too severe or too lenient would impact the validity and reliability of the assessment. By conducting a Differential Rater Functioning (DRF) analysis using MFRM on a middle school 7th-grade writing skills assessment in Turkey, they tried to identify the potential unfairness judgment in that assessment (Aslanoğlu & Şata, 2021). The results showed that MFRM could identify elements in the rubric that functioned differently or discriminated against certain groups (Aslanoğlu & Şata, 2021). Similar research also conducted by Arifiyanti et al. (2023) by investigating the interaction between raters and students in the assessment of research seminar presentations. Through MFRM, they able to identify bias patterns caused by gender and student academic majors (Arifiyanti et al., 2023). In the context of art-based competitions, Nor et al. (2024) demonstrated MFRM's ability to analyse the quality of the rubrics and assess the fairness of competitions organized by the Creative Wood Workers Group (GSKK). This study recommends MFRM because its capacity to provide bias/interaction reports increases its usefulness, particularly in the context of competitions (Nor et al., 2024). This is crucial because the subjectivity in the assessment, particularly of art-based performance that evaluate aesthetics, are susceptible to biased results (Brooker & Antonini, 2025; Scherbakova et al., 2025).

Many researchers specifically use MFRM to see the rater effect because each of whom has the potential to have subjectivity. Kasim (2011) explored judging behaviour and rater errors that demonstrates rater severity, restriction of range, internal consistency, and central tendency. Cai (2015) conducted a weight-based classification of raters and rater cognition in the EFL speaking test. Springer et al. (2018) investigated adjudicator bias in concert band evaluation. Styck et al. (2020) evaluated rater effects in classroom assessment scoring system. Wang et al. (2020) also looked at rater performance on the CELBAN speaking. Han (2021) conducted a test by comparing classical test theory, generalizability theory, and MFRM to see the ability of the three to measure rater effects. Mohamat et al.'s (2022) study examined raters' assessment quality in measuring teachers' competency in

classroom assessment. Wang and Long (2022) used MFRM to reexamine the subjective creativity assessments in science tasks by considering the rater-mediated assessment framework. All studies on rater effects were conducted using MFRM and recommends on the need for training in assessment, especially for novice raters, were widely voiced.

Previous research has shown that applying MFRM provides many benefits, especially when assessing the quality of an instrument and an assessment involving many raters. This model of analysis is highly flexible in terms of psychometrics and rater assessment patterns, which can detect indications of bias that affect the assessment's accuracy. Its application is also very diverse, from cognitive to non-cognitive, such as performance or art-based assessment. As a valuable reference, evaluating the Quranic recitation rubric assessment is possible, especially if the intention is to evaluate its psychometric aspects. Moreover, research on MTQ assessments that utilize MFRM as a modern analysis tool is still difficult to find.

2.1.5. Challenges and Gaps in Current Research

One of the significant but not addressed limitations of the MTQ competition is the lack discussion regarding the issues on the assessment of Quranic recitation made by judges. A review of the extant literature reveals that Quranic recitation competitions, play a significant role in the educational realm. Beyond its function as an informal educational program, the MTQ competition has exerted a profound influence on educational institutions and wider society. However, despite the positive impact, some studies have identified controversies that often arise due to the MTQ assessment system. These controversies encompass a range of issues, including the perceived lack of transparency in the assessment process, allegations of politicization in the evaluation outcomes, and the credibility and reliability of the assessment rubric, particularly in terms of addressing potential biases among assessors. Some studies attribute these concerns to the inadequacy of assessment practices in aligning with the principles of measurement and assessment.

Research on the MTQ assessment issue has predominantly adopted a qualitative approach, which has yielded valuable insights but cannot provide comprehensive answer. Unfortunately, there are very few studies that utilize a quantitative approach. In fact, this method will be very valuable, especially in providing empirical data in the case of evaluation of the assessment rubric. In fact, many studies have utilized advanced analysis tools, namely MFRM, to provide valuable information about the quality of assessments, from the psychometric aspects to the systematic pattern of assessment given by the rater. Unfortunately, the complexity of the Quranic recitation assessment rubrics that require

MFRM analysis are difficult to find. Even though Bahruddin and Khumaidi (2014) research has developed Quranic recitation assessment rubric by identifying causative factors, a literature review method was used instead of MFRM, which is a more suitable method for evaluating the quality of the rubric.

For those reasons, it can be synthesized that the validity, reliability, and fairness of the Quranic recitation assessment rubric in the MTQ competition have not yet been tested using MFRM as an analytical tool. The present study aims to address this research gap by assessing the validity, reliability, and fairness of the Quranic recitation assessment rubric in the MTQ competition. It is hope that the assessment rubric, which has been the catalyst for the numerous issues and controversies that have emerged in the MTQ competition, can find its position as a valid, reliable, and fair assessment rubric or vice versa.

2.2. Theoretical Framework

2.2.1. Rasch Model Theory and Its Extension to Many-Facet Rasch Measurement

In this study, the Many-Facet Rasch Measurement model developed by Linacre (1989) was utilized. This model is an extension of the Rasch model developed by George Rasch (1960). To provide a comprehensive understanding of the rationale behind this selection, a detailed narrative will be presented elucidating the model's emergence and capabilities, as well as its application in this study.

Initially, assessment in education was regarded as a nascent discipline. During the 19th century, the practice of assessment emerged as a new discipline and began to be incorporated into formal education. Historically, the practice of assessment, as embodied by examinations, traces its origins to 260 BC during the Han Dynasty in China, where it was utilized for the selection of qualified royal bureaucrats (Sumintono & Widhiarso, 2015). During the 16th century, this examination system was introduced to Europe by European missionaries and adapted for the selection of employees in both government and private spheres, such as the English East India Company (Sumintono & Widhiarso, 2015).

In the 1800s, the practice of exams, as they are now known, began to be implemented. The demands for educational reform during the Industrial Revolution encouraged the use of exams to assess the achievements of institutions in Europe (Sumintono & Widhiarso, 2015). Concurrently, in 1845, the American community, particularly in Boston, began transitioning from oral to written examinations (Marzona, 2000). According to Brookhart (2016; 2019), this transition was driven by the desire to

facilitate a comparative analysis of educational achievements across different schools, with the goal of providing feedback and enhancing motivation.

The use of tests as an evaluation tool requires the use of analytical instruments to ensure the effectiveness of the assessment instrument. In the 20th century (1904), Charles Spearman proposed a measurement theory known as the Classical Test Theory (CTT) (Traub, 1997). This theory enables the prediction of test results by considering various parameters, including student ability and item difficulty (Andrich, 2019). The basic assumption is that a score consists of a pure score and measurement error (Bond et al., 2021). However, the use of raw scores in CTT theory has several weaknesses, such as raw scores are not measurement results, have weak quantitative meaning, do not indicate a person's ability to perform certain tasks, and raw scores and percentages of correct answers are not always linear (Sumintono & Widhiarso, 2015). Due to these weaknesses, CTT theory has begun to be widely abandoned (Lord & Novick, 1968).

In the 1950s, Item Response Theory (IRT) began to be developed to overcome the weaknesses of CTT. IRT models the relationship between individual abilities and item characteristics probabilistically (Hambleton et al., 1991). IRT has three types of measurements; namely the 1PL model involving one parameter in the form of the level of difficulty of the test items, the 2PL model involving one parameter in the form of the level of difficulty and discrimination power of the test items, the 3PL model involving parameters in the form of the level of difficulty, discrimination power, and pseudo guesses of the test items (Sumintono & Widhiarso, 2015). In the following decade, a new model emerged that had similarities with the IRT 1PL model because it emphasized the level of difficulty of the test items, however, both have differences in their measurement models (Van der Linden, 2016; Engelhard & Wang, 2021). This model was later known as the Rasch model (Boone et al., 2014; Sumintono & Widhiarso, 2015; Van der Linden, 2016).

George Rasch originated the Rasch model in the 1960s, a special form of IRT. The Rasch model is designed to measure the probabilistic relationship between test takers' abilities and the difficulty of items (Rasch, 1960), producing linear measures that are invariant to samples and items (Bond & Fox, 2015). As an invariant parameter, the Rasch model can independently measure the difficulty level of an item and the ability of a participant. Due to its simple yet powerful nature, the Rasch model provides the solution for objective measurement (so far), particularly in the context of the educational field (Wright & Stone, 1979).

In this modern era, the Rasch model is useful not only for dichotomous data but also for polytomous data. David Andrich developed the Rasch Rating Scale Model (RSM), an alternative Rasch model analysis, to accommodate polytomous data (Andrich, 2019). Other alternatives of analysis have also been developed to accommodate the involvement of multiple raters in the practices of measurement. Linacre (1989) introduced an innovation called Many-Facet Rasch Measurement (MFRM), a branch of the Rasch model intended to overcome previous limitations. By incorporating additional facets such as rater, task, or item, and demographical backgrounds, MFRM is intended to accommodate a complex assessment context (Eckes, 2011; Linacre, 2023).

According to Linacre (1989), the MFRM differs from the original Rasch model, yet it still maintains the basic principle of the Rasch model, producing linear dimensions by involving many dimensions in the analysis process. There are several advantages if the assessment evaluation is decided using MFRM. First, it can identify assessor bias (Eckes, 2011). Second, MFRM can overcome variation in item difficulty (Bond & Fox, 2015). Third, MFRM ensures that the measurement results are more accurate, valid, and reliable (Linacre, 1989). MFRM can also measure assessor consistency, increasing confidence due to its detailed and comprehensive results (Wright & Linacre, 1989). MFRM can be applied to performance-based assessments in a variety of fields (Box & Fox, 2015; Eckes, 2011; Linacre, 1989).

In the context of this study, MFRM will be used to assess the quality of the Quranic recitation rubric assessment in the MTQ competition. MFRM was selected because it can accommodate multi-rater assessments, reveal psychometric properties, explain rater and ratee performance and item difficulty distribution, and detect assessment bias affecting the quality of assessment. Following the format of the Quranic recitation rubric, the MTQ guidebook explains that some items are not directly equivalent to the scores for other items; therefore, different types of data will be produced. To address this challenge, specific data coding procedures that MFRM can accommodate will be implemented.

2.2.2. Framework of Validity, Reliability, and Fairness

The objective of this study is to assess the Quranic recitation assessment rubric. As a measurement tool for assessment, Engelhard (2013) explained that several measurement principles must be fulfilled. Referring to The Standards for Educational and Psychological Testing, three foundational areas must be considered when evaluating measurement procedures: validity, reliability, and fairness. These foundational areas are a set of professional guidelines developed by a joint committee of the American Educational

Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) in 2014. These guidelines are standards commonly used to develop, evaluate, and implement tests, particularly in psychology and education, ethically and effectively. They include testing the technical quality of psychometric tools in assessments. Since this study aims to evaluate the technical aspects (psychometrics), these standards were selected. Consequently, three foundational areas will be detailed, as they form the basis of the theoretical framework in this study. The three foundational areas will consistently be referred to as the "*Standards*."

The first standard is validity. According to the consensus view of validity from the Joint Committee is as follows:

> *"Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations."* (The standards, 2014, p. 11)

Referring to the standard 1.0, the essential point of validity is "Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided." This concept is also confirmed by general researchers such as Messick (1989), who defines validity as a measuring instrument that can measure something that should be measured; in other words, validity is about the argument, then findings evidence for that. Validity is a continuous process of proof, referring to the extent to which evidence and theories support the interpretation of test scores according to the purpose of the test, building a validity argument (Aryadoust, 2013; Sumintono & Widhiarso, 2015). Engelhard (2013) asserts that, within the context of measurement, validity is closely associated with the content of the test, empirical evidence, and conceptual arguments that support the interpretation of measurement results.

It should be noted that this study is using the MFRM model, which needs to be considered in determining the indicator used to determine validity of the Quranic recitation assessment rubric. The *Standards* define at least three thematic clusters in Standard 1.0, namely: establishing intended uses and interpretations, issues regarding samples and settings used in validation, and specific forms of validity evidence. Each of cluster have

the source of evidence that can be used to interpret the validity of the data. To provide a clear picture, Tabel 2.1. presented as follows:

Table 2. 1 *The source of validity evidence*

No	Cluster		Source of evidence
1	"*Establishing intended uses and interpretations*."		Definition of the construct Description of the scale Wright map (intended/hypothesized)
2	"*Issues regarding samples and settings used in validation*."		Detailed description of the person samples and observation settings
3	"*Specific forms of validity evidence*."	Test content	Expert judgment Content validity studies Empirical item hierarchy Wright map (empirical)
		Response processes	Scoring rules Model-data fit (person fit) Person response functions Cognitive studies using qualitative methods
		Internal structure	Wright map Model-data fit (items, persons) Evaluation of ordering of items and persons Unidimensionality
		Relations to other variables	Program of research on the scale Differential item functioning across subgroups of persons Differential person functioning across subsets of items Criterion-related evidence
		Consequences test	Program of the research on the scale

Cited with some adjustments from the Standards (2014) and Engelhard & Wang (2021)

Table 2.1 lists the three clusters from the Standards (2014) and some adjustments from the book of Engelhard & Wang (2021) to describe the source of evidence. In cluster

number one, the validity measured focuses on establishing the intended uses and interpretations of the scale scores. Commonly in education, the intended uses include providing information on student performance from the summative or formative assessment that has been conducted (Engelhard & Wang, 2021). In this study, the intended uses of the scale scores focus on measuring the ability of students to recite the Quran properly and beautifully. In terms of the evidence, this study can explore the definition of the construct, description of the scale, and the intended or hypothesized Wright map (The Standards, 2014). In cluster number two, the focus of the evidence will be on the issues related to the samples (persons), including the settings of the assessment. The aim is to guide whether the rubric used in the MTQ competition is appropriate for the intended groups and settings. For that reason, a detailed description of the persons as samples and observation settings is needed.

For cluster number three, validity is intended to measure the content, constructs, and criteria. Specific forms of validity evidence can be gained by analysing five forms of validity, including: the test content, response processes, internal structure, relations to other variables, and consequences testing. *Test content* explains the relationship between the content of the test and the construct being measured (The Standards, 2014). Several sources of evidence can be used, including expert judgment to support the meaning of the item test, content validity studies, and the Wright map to evaluate the item hierarchy ordered from easy to endorse, particularly in providing empirical evidence regarding the meaning of scale scores on the latent variable. This evidence can also be developed through logical or empirical analysis of the test's content domain, the theoretical foundation utilized, the language employed in the question items, and other test properties (Sumintono & Widhiarso, 2015).

Response processes explain how an individual interacts with the items on the scales. Evidence based on the responses provided by the test participants, particularly if this process is relevant to the measured answers, indicates the validity of the tests and items (Engelhard, 2013; Wind, 2023). This is crucial for evaluating the function of rating scales (Engelhard & Wind, 2018). The source of its evidence can be collected from scoring rules, person fit analysis within a model (especially if the person’s response is aberrant), and cognitive studies utilizing qualitative methods.

Internal structure explains the extent to which test items and scale categories conform to a hypothesized Rasch scale (The Standards, 2014; Wind, 2023). Various statistical analyses can unveil internal structures, such as the Wright map, to show the order

of the items and persons on the same scale. The Wright map can be examined by analysing the model-data fit of persons and items to determine whether the internal structure of the scale supports the inference as intended (The Standards, 2014). The examination of unidimensionality and evaluation of persons and items ordering can also provide meaningful information about the internal structure of the rubric.

The next source of evidence comes from the *Relations to other variables*. Evidence based on association with other variables, including evidence of convergence and discrimination (The Standards, 2014). Convergence highlights the relationship between developed test scores and other test scores that measure the same construct. Discrimination focuses on the relationship between the test scores compiled and other tests that measure different constructs (Wesolowski, 2019). Determining whether the items and the scales in the rubric relate to other substantive variables is crucial. According to Engelhard & Wang (2021), this is important for building up broader theoretical support for the rubric. Some source of evidence, such as differential item functioning and differential person functioning across persons and items, is frequently used to interpret the validity of the rubric. Criterion-related evidence can also support the validity of the rubric.

The final evidence in cluster three is the *Consequences of testing.* The development of the rubric is always used to serve a specific purpose. In the context of test development, it is important to consider the intended and unintended consequences so that the profound effects can be anticipated. Evidence based on test consequences is considered valid if the consequences are positive and the test instrument is deemed valid. Conversely, if the results are adverse, the test instrument has a low level of validity (The Standards, 2014). Relating to the use of the Quranic recitation rubric, which is utilized by so many different contexts (groups, times, assignments), examining the consequences of the assessment is important.

The second standard is reliability. According to the Standards (2014), the concept of reliability can be defined as follows:

> *"... the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form. ...in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported."* (The standards, 2014, p. 33)

Referring to the standard 2.0, the essential point of reliability is "Appropriate evidence of reliability/precision should be provided for the interpretation for each intended

score use." Cronbach (1951) also stated that reliability indicates the consistency of measurement results that, when replicated, persist within analogous parameters. In the Rasch framework, reliability is assessed not only through conventional indices but also through separation indices, which facilitate the mapping or differentiation of individuals and items (Engelhard, 2013). The concept of reliability is summarized by three different terms: stability, equivalence, and internal consistency (Sumintono & Widhiarso, 2015). Stability indicates the occurrence of consistent results when a test is administered repeatedly. Equivalence occurs when identical results are produced by two tests when compared. Internal consistency relates to the uniformity of the score results of each item in a test. For this reason, in measuring reliability, a comparator is needed. Comparison between time or retest, comparison between forms of test or parallel test, and comparison between test components is an internal consistency approach (Wilson, 2005; Engelhard & Wang, 2021).

The consensus of Joint committee describes the reliability within eight interrelated clusters. These clusters can be grouped into three topics: research, indicators, and documentation (Engelhard & Wang, 2021). The group of research is related with the evidence regarding the consistency and precision of person measures across various test replications. The group of indicators are functioned to determine which indicator can be used for evaluating the consistency and precision of the rubric. The group of documentation is intended to provide the support evidence for other clusters which can be collected from the other publications. All the interrelated clusters are shown in the Table 2.2.

Table 2. 2 *The source of reliability evidence*

Topic	**No** *(not in order)*	**Cluster**	**Source of evidence**
Research	1	"*Specifications for replications*."	Program of research on the scale
	2	*"Evaluating reliability and precision."*	Program of research on the scale
	4	*"Factors affecting reliability and precision."*	Program of research on the scale

	7	*"Reliability and precision of group means."*	Program of research on the scale
Indicators	3	*"Reliability and generalizability coefficients."*	Reliability of separation for both persons and items
	5	*"Standard errors of measurement."*	Standard errors of measurement for persons Standard errors of measurement for items
	6	*"Decision consistency"*	Indices of decision consistency
Documentation	8	*"Documenting reliability and precision."*	Documenting the invariance of scale scores

Cited with some adjustments from the Standards (2014) and Engelhard & Wang (2021)

Table 2.2 lists the eight clusters from the Standards (2014), which are grouped into three interrelated topics. Considering the source of evidence, cluster numbers 1, 2, 4, 7, and 8 provide the evidence from the program of research on the rubric to understand its invariance. By exploring whether the rubric is achieved through the invariant measurement, the researchers can seek the aspect of assessment situation that yields variance measures on the rubric. This process can be viewed as part of evaluating the consistency and invariance of person scores.

For clusters number 3, 5, and 6, the Standards (2014) mentioned that the evidence gathered from the psychometric indicators can be used for reporting the consistency and precision of the rubric over various replications. In the Rasch measurement theory, the consistency and precision evidence can be viewed by considering several interrelated indicators, including: The reliability of person scores, the precision of measures, the reliability of item scores, and the precision of item calibrations (the Standards, 2014). According to Cronbach (1951), as cited in Engelhard & Wang (2021), the reliability of a person can be assessed by the coefficient alpha, while precision can be interpreted from the standard error of measurement report. Both aspects were useful for explaining person variability, the spread of items in the rubrics, and the precision of the calibration of the

items. In the context of this study, the evidence that supports revealing the rubric quality in terms of reliability will be used.

The third standard is fairness. According to the Standards (2014), the concept of fairness is relied on the understanding that the fairness is *"a fundamental validity issue and requires attention throughout all stages of test development and use. Fairness to all individuals in the intended population of test takers is an overriding, foundational concern.*" (The standards, 2014, p. 49). Specifically, to the standard 3.0, the essential point of fairness is "*All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.*" (The standards, 2014, p. 63). Some interrelated clusters were described by the standards following the source of evidence that can be used to evaluate the rubric of assessment. These clusters are presented in Table 2.3.

Table 2. 3 *The source of fairness evidence*

No	Cluster	Source of evidence
1	*"Test design, development, administration, and scoring procedures that minimize barriers to valid score interpretations for the widest possible range of individuals and relevant subgroups."*	Program of research on the scale Differential item functioning
2	*"Validity of test score interpretations for intended uses for the intended examinee population."*	Differential person functioning
3	*"Accommodations to remove construct-irrelevant barriers and support valid interpretations of scale scores for their intended uses."*	Program of research on the scale Identification of construct-irrelevant barriers
4	*"Safeguards against inappropriate score interpretations for intended uses."*	Program of research on the scale

Cited with some adjustments from the Standards (2014) and Engelhard & Wang (2021)

The first cluster of fairness is focused on searching for the irrelevant variance in the construct of the rubric. The second cluster also stresses the importance of ensuring the fairness of interpretation to the person who used the rubric. By analysing and identifying the problem in the construct, a researcher can provide evidence on which items or scales function differently. By analysing the construct of the scoring rules, a researcher can capture the variance of each individual or subgroup, such as gender or ethnicity (Engelhard & Wang, 2021). The source of evidence suggested by the Standards (2014) can be found by considering the report on the model-data fit, including differential item and person functioning. In the context of Rasch analysis, this evidence can define the latent variable that explains differential meaning for items, scales, individuals, or subgroups. The third and fourth clusters emphasized evaluating the rubric to consider how to address construct-irrelevant barriers. Some issues discovered need to be addressed through a careful research process on the rubric to determine which aspects should be revised or changed. The analysis of model-data fit, differential item functioning, and differential person functioning will support the recommendations for creating a fair rubric construct. The Standards (2014) pointed out that "*A test that is fair… reflects the same construct (s) for all test takers, and scores from it have the same meaning for all individuals*" (The Standards, 2014, p. 50).

All the clusters with their source of evidence explained in each foundational area are the warrant standards constructed by the AERA, APA, and NCME. Every analysis of the validity, reliability, and fairness is intended for its purpose, which the focus of the study may not require all the evidence listed here. All the psychometric reports on the rubric evaluations are about the argument from the evidence gathered. It also suggests that the weight placed on each source of evidence may vary across purposes (Engelhard & Wang, 2021). Therefore, for that rationale, the use of evidence in this study will specifically choose the evidence that can support the interpretation of the rubric qualities' evaluation.

To examine the validity of the rubric, the evidence presented in Table 2.1 will be collected from the definition of the construct, description of the scale, detailed description of the samples and settings. The other evidence from the Wright map, model-data fit, items ordering, unidimensionality in the internal structure, and the scoring rules dan person response functions in the response process will also be considered. For the reliability, all the evidence presented in Table 2.2 related to the indicator's topic, which is reliability of items and persons, separation of items and persons, indices of consistency, and all the precision indicators, such as the standard error measurement, will be collected. Furthermore, the evidence collected to interpret the fairness will rely on the report from the

model-data fit and differential item and person functioning as presented in Table 2.3. All the evidence of the three foundational areas above is collected separately. However, the Standards (2014) emphasized that all evidence from three areas functions together. Therefore, in evaluating the Quranic recitation rubrics, this study will embed the discussion by involving all the results.

In the context of this study, the testing of the quality of the Quranic recitation assessment rubric refers to its validity, reliability, and fairness. Validity testing determines whether the rubric measures what it should measure according to the underlying theory. Reliability testing ensures the consistency of the rubric when used by different individuals at different times. Similarly, the fairness test will examine the bias between facets of the Quranic recitation assessment rubrics in the MTQ competition.

2.3. Conceptual Framework

This section explains the conceptual framework that will be used in this research. This framework is created from the integration of two different concepts: the assessment of the Quranic recitation rubric and the MFRM analysis model. The Quranic recitation rubric assessment framework is based on the specific rules of the MTQ competition. The MFRM framework also followed the principles underlying the MFRM model. The conceptual framework developed from these two concepts is illustrated in Figure 2.1 and will serve as the main framework to guide the research.

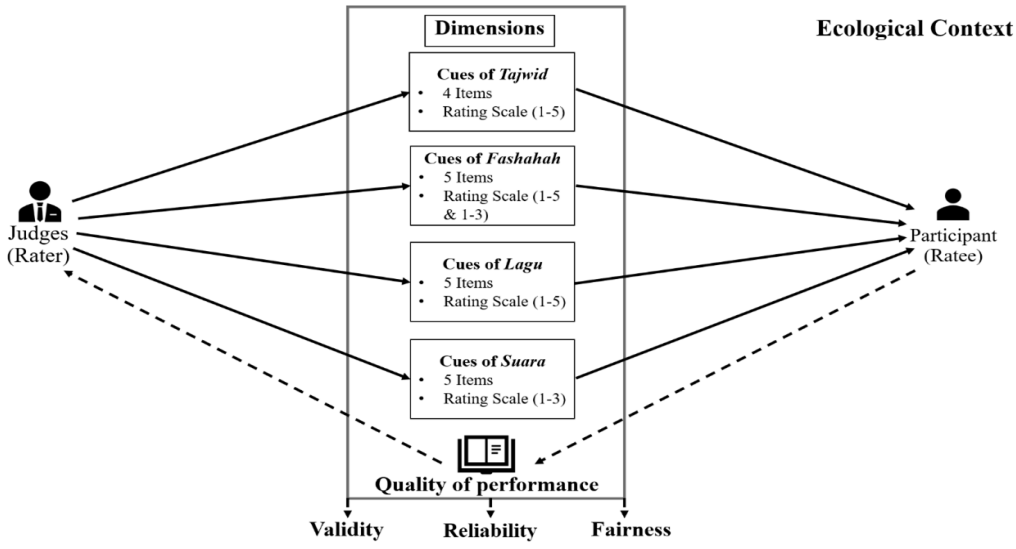


Figure 2. 1 *Conceptual framework*

In general, this conceptual framework is guided by the MFRM analytical framework, built based on Brunswik's Lens Model (1952) to understand human perception and judgment. The figure shows that there is an interconnected relationship between three components, which will be referred to as “facets.” One of these facets is the judges, who serve as raters who provide assessments based on the rubric that has been provided. The assessment rubric, as defined by Cooskey (1996), is a series of dimensions referred to as “cues.” Each cue consists of several items with certain rating scale conditions (Cooskey, 1996). These components, in essence, constitute an assessment rubric. According to Landy and Farr (1980), these rubrics should be utilized to evaluate the performance of participants. After the collection of assessment results for each participant, the dataset can be utilized for the evaluation of various aspects of the assessment process, including the distribution of items and participant performance, the quality of the judges' assessments, and the quality of the items in the assessment rubric employed. This concept aims to measure invariants with an appraiser through a series of cues that have been determined (Hogart, 1987). The following five critical requirements must be met: “rater-invariant person measurement, rater-invariant calibrations of cues and rating scales, invariant locations of raters, and rater-invariant Wright maps” (Engelhard & Wind, 2018, p. 262). It is imperative to note that all these steps occur within a particular ecological context and are to be utilized according to the requirements of the research objective, which is to assess the validity, reliability, and fairness of the Quranic recitation assessment rubric in the MTQ competition.

The assessment rubric that will be evaluated is the Quranic Recitation Rubric Assessment, which is analogous to the assessment concept employed in MTQ competitions in Indonesia. According to the "Musabaqah Guidebook of Al-Quran and Al-Hadith" (2023), the Quranic recitation assessment rubric encompasses the same facet components as the MFRM concept, comprising rater, assessment rubric, and participants. The rater, in this case, consists of judges comprising 12 individuals, all of whom will be involved in the assessment of each participant. Each judge will be divided into four groups, with each group assigned to assess one of the four dimensions. The composition of each group will be three judges, interacting in a manner to ensure independence in their assessments (Engelhard, 2013). The dimensions incorporated in the rubric encompass *Tajwed, Fashahah, Lagu,* and *Suara*, with each dimension comprising 4-5 items. These items serve as guidelines for the judges, ensuring the quality of participants' performance from the perspectives of practical theory and aesthetics. The scoring criteria for each item vary, contingent on the quality and errors of the participant's performance. To streamline the data processing procedure, a likert scale will be formulated to establish standardization. The

details of this explanation will be discussed further in the methodology chapter. Apart from that, it should be acknowledged that there are limitations in this study. The other possible facets related with this assessment framework such as the assignment for the participant's recitation (*maqra')* or the time setting of the assessment are not included as the facets in data analysis.

CHAPTER III
METHODOLOGY

The methodology section in this chapter provides a comprehensive overview of the technical aspects underlying the methodological implementation of the study. The philosophical theory behind this study, the approach employed, the research design used, the method chosen, the research subject, the data collection techniques and procedures, the research instruments utilized, the data analysis procedures, and the ethical considerations will be explained. This systematic explanation is used to offer the comprehensively construct of the methodology.

3.1. Theoretical Philosophy Underpins the Methodology

This research is based on the theoretical foundation and relevant paradigms to understand the quality of Quranic recitation assessment rubrics in the MTQ competition. Theoretically, this research is rooted in the positivist paradigm, which, referring to Cohen et al. (2018), is a philosophical approach based on the assumption that reality is objective, observable, and empirically measured using scientific methods. The positivism paradigm encourages a researcher to use of quantitative methodology, which usually involves statistical measurements and provides generalization of its findings (Creswell & Creswell, 2018). According to nature this paradigm, a researcher is considered as a neutral and does not give any effect to the results of research as they prioritize the nature of quantitative methods, prioritize structured measurements (Bryman, 2016).

In the context of this study, the positivist paradigm was very relevant for the objectives of assessing the quality of Quranic recitation assessment rubrics. Since its given emphasis on objectivity, empirical measurement, and generalization of results (Ali, 2024; Pallant, 2016). Under its capacity, the measurement of validity, reliability, and fairness can be empirically and systematically carried out by utilizing the numerical data obtained (Creswell & Creswell, 2018). The use of this paradigm has ensured that the results obtained provide strong and reliable empirical evidence for the evaluation and development of Quranic recitation assessment rubrics in the MTQ competition.

3.2. Research Approach

In line with the research paradigm, the quantitative approach was chosen because its philosophical underpins and its design which able to accommodate the numerical data utilizing a systematic procedure to draw generalizable conclusions (Creswell & Creswell,

2018). This selection was intended because of this research focused on assessing the psychometric attributes: the validity, reliability, and fairness of Quranic recitation assessment rubrics by using the numerical data obtained. By adopting the national Quranic recitation rubric, this study was expected to provide generalizable findings, as suggested by Creswell and Creswell (2018). However, it must be noted that the results depend on the research properties, such as the sampling technique chosen.

3.3. Research Design

The instrument testing design is utilized in this study because the validity, reliability, and fairness of the Quranic recitation assessment rubric in the MTQ competition is aimed to be tested. According to The Standards (2014), validity testing determines how accurately an instrument captures the concept it is designed to measure. Reliability testing aims to assess the consistency of measurements over time and among assessors (Fraenkel et al., 2012). The test will also reveal how fairly the assessment rubric is produced. This design aligns with the objectives and contexts of the study, which aims to determine if the Quranic recitation assessment rubric in the MTQ competition adheres to the principles of measurement and assessment (Bond & Fox, 2015).

It should be emphasized that this study is non-experimental because the researcher did not manipulate or intervene in any aspect (Johnson & Christensen, 2014). Instead, the researcher only observed the phenomenon as it occurred naturally (Carlson & Morrison, 2009). However, in terms of data collection specifically, the researcher designed activities that were as similar as possible to the MTQ competition practice. This replication was carried out due to time constraints, limited access, and costs to create or participate in real competition activities that occurred in society. For this reason, the study was adapting and replicating the competition at a certain period so that the results given cannot be used to determine a definite cause-and-effect relationship, because it was considered a cross-sectional method (Levin, 2006).

3.4. Research Subject

This study used the convenience sampling method to recruit participants as research subjects. According to Creswell and Creswell (2018), this sampling technique involves selecting subjects based on their availability, ease of access, and willingness to participate. The selection was made considering the time, cost, and resource limitations that hindered the application of more complex sampling methods. While this method is helpful in terms of efficiency and ease of implementation, it is important to acknowledge its drawbacks

regarding generalization because the sample was not fully representative of the target population. To overcome this limitation, the study tried to maximize the involvement of the population in the selected sample.

The research location chosen for the case study was an Islamic educational institution, or Islamic boarding school (*Pesantren*) in the province of Banten. This school was selected because its students regularly participate in the MTQ competition, particularly in the Quranic recitation competition, at various levels every year. The *Pesantren* also claimed its status as a "Center for Quran Recitation Studies".

The study participants were teachers, *Pesantren* officers, alumni, and students, who were familiar with the competition. Initially, the participants were raters who had experience as judges, participants, pedagogues (coaches) in MTQ competitions, and or had participated in specialized training as MTQ judges, which was provided by the Tilawatil Quran Development Institute (LPTQ). Sixteen participants who fulfilled the judge's requirements were selected as judges (Raters). Since experts in Qur'anic recitation were limited at this *Pesantren*, teachers who specialized in Quranic fields, such as *tahsin* and *tajwed,* were also involved. Due to the limitations of their specialization, particularly in the *nagham* science (the science of beautifying Quranic recitation), these additional judges focused on aspects of theoretical and technical aspects, such as *Tajwed* and *Fashahah*, rather than on the aesthetic aspects, such as *Lagu* and *Suara*. The evaluation of aesthetic and vocal aspects was conducted exclusively by judges who are experts in the practice of Quranic recitation and *nagham* science. This delineation of roles was intended to ensure objectivity and accuracy in the assessment process (Gibbons, 2015; Bond & Fox, 2015). Before the assessment process began, all participating judges attended workshops or training sessions designed to ensure a uniform understanding of the assessment rubric. The training materials referred to the official guidelines of the MTQ, and the workshops were led by expert speakers who have demonstrated competence and experience at the national MTQ competition level.

The other participants as performers in this study were 50 students who enrolled in the *Tilawah* Al-Quran class at this Islamic boarding school. All students were involved based on their shared background of receiving education and training in Quranic recitation, including *tahsin, tajwed*, and *tausyikh* (*Nagham* science). Related to that, the demographic information of the judges and students as performers is provided in Tables 3.1 and 3.2:

Table 3. 1 *The demographic information of judges*

Judges Demographics	**Type**	**Frequency (n: 16)**	**Percentage (%)**
Gender	Male	6	37.5%
	Female	10	62.5%
Origin of province	Sumatera Utara	1	6.25%
	Banten	2	12.5%
	DK Jakarta	1	6.25%
	Jawa Barat	1	6.25%
	Jawa Timur	3	12.5%
	Bali	1	6.25%
	Nusa Tenggara Barat	1	6.25%
	Sulawesi Utara	3	18.75%
	Sulawesi Selatan	1	6.25%
	Gorontalo	2	12.5%
Age	18	1	6.25%
	19	3	18.75%
	20	3	18.75%
	21	3	18.75%
	23	2	12.5%
	24	1	6.25%
	25	1	6.25%
	26	1	6.25%
	29	1	6.25%
Long stay in the Pesantren (years)	1 year	2	12.5%
	2 years	1	6.25%
	3 years	3	18.75%
	4 years	1	6.25%
	5 years	1	6.25%
	6 years	3	18.75%
	7 years	1	6.25%
	8 years	1	6.25%
	9 years	2	12.5%
	11 years	1	6.25%

Positions	Teacher	1	6.25%
	Alumni	5	31.25%
	Undergraduate student	1	6.25%
	Teacher and officer	9	56.25%
The field of experience in the MTQ	1 field (Participant)	4	25.0%
	2 fields (Participant and judge or Participant and pedagogue)	10	62.5%
	3 or more fields (Participant, judge, and pedagogue)	2	12.5%
Expertise in MTQ	1 expertise (*Tilawah* or *Tahfidz*)	5	31.25%
	2 expertise (*Murottal and Qiraat* or *Tahfidz and Tafsir* or *Tilawah and tahfidz* or *Tilawah and Syarhil*)	7	43.75%
	3 expertise (*Tilawah, Tahfidz, and Qiraat* or *Tilawah, Murottal, and Qiraat*)	2	12.5%
	4 or more expertise (*Tilawah, Tahfidz, Syarhil, and Fahmil* or *Tilawah, Tahfidz, Syarhil, Fahmil, Murottal, and Qiraat*)	2	12.5%

Table 3. 2 *The demographic information of performers*

Performers' Demographics	**Type**	**Frequency (n: 50)**	**Percentage (%)**
Gender	Male	29	58.0%
	Female	21	42.0%
Origin	Sumatera Utara	2	4.0%
	Bengkulu	3	6.0%
	Jambi	3	6.0%
	Lampung	1	2.0%
	Sumatera Selatan	2	4.0%
	Banten	11	22.0%
	DK Jakarta	6	12.0%

	Jawa Barat	4	8.0%
	Jawa Tengah	1	2.0%
	Jawa Timur	2	4.0%
	Kalimantan Barat	1	2.0%
	Kalimantan Tengah	3	6.0%
	Bali	1	2.0%
	Nusa Tenggara Barat	2	4.0%
	Sulawesi Utara	2	4.0%
	Sulawesi Tenggara	1	2.0%
	Gorontalo	3	6.0%
Age	Junior level (< 15)	19	38.0%
	Senior level (≥ 15)	31	62.0%
Grade level	7	9	18.0%
	8	5	10.0%
	9	7	14.0%
	10	13	26.0%
	11	7	14.0%
	12	7	14.0%
Long stay in the	1 year	18	36.0%
Pesantren (years)	2 years	10	20%
	3 years	9	18.0%
	4 years	9	18.0%
	6 years	4	8.0%
Experience in	1-5 times	17	34.0%
MTQ	6-10 times	7	14.0%
	>10 times	14	28.0%
	None	12	24.0%

3.5. Data Collection Technique and Procedure

This research employed a set of data collection techniques, all of which were subject to rigorous procedural requirements. The main technique was the completion of a questionnaire (Johnson & Christensen, 2014). The questionnaire was adopted based on a Quranic recitation assessment rubric instrument from the MTQ competition. The details will be explained in the subsequent sub-section entitled "Research Instrument." All

processes were executed according to the MTQ competition's assessment standards, a point accentuated through workshop or training for the judges. To support the validity of the collected data, supporting documentation such as photos and videos, and written observations are also provided (Creswell & Creswell, 2018).

In the process of data collection, the researcher carried out several stages, all of which were completed to ensure systematic data collection (Appendix 1). First, official permission was obtained from the *Pesantren*, as well as the official permission and research recommendation from LPTQ national and LPTQ Banten province (Appendix 2). Second, the researcher recruited one main research assistant and several accompanying research assistants, who received training and orientation related to data collection based on the created research protocol (Appendix 3). Research assistants were recruited to help manage the research intervention, particularly the recruitment of participants and the collection of their consent and demographic data. The research assistants are *Pesantren* officers who understand the context of the *Pesantren* better.

Third, the recruitment of participants successfully included 16 raters (judges) and 50 performers who were selected based on their willingness, qualifications, and demographic data information (Appendix 4). Fourth was the implementation of an online workshop on filling out the assessment rubric, guided by experienced national judges from Bali, Indonesia, as well as the distribution of related learning materials (Appendix 5). Fifth, the assessment was implemented and carried out for one full day on March 1, 2025 (Appendix 6). Sixth, assessment data was collected. In this study, data were obtained from 16 raters who assessed 50 participants on 19 items across four dimensions, producing 3760 quantitative responses and 674 qualitative responses. Ideally, 3800 quantitative responses and 800 qualitative responses would have been collected. However, there are 40 quantitative data responses are missing and 126 data sheets with no qualitative responses. It is important to note that qualitative data is only collected as a record of the responses made by the raters (judges), considering that the rubric format is adopted as similar as possible to the rubric used in the national competition. Therefore, this data collection is not intended to fulfil the requirements of other research methods such as qualitative or mixed methods. This data only reports the qualitative responses given by the raters (judges), which will only be used as supporting data for the findings of the quantitative data analysis results (if necessary).

3.6. Research Instrument

The assessment instrument used in this study is the Quranic Recitation Assessment Rubric, which was adopted from "Guidelines for Musabaqah Al-Quran and Hadith in 2023" (2023). Use of the rubric has been permitted ethically by LPTQ national, the owner of the rubric. This rubric consists of four dimensions: *Tajwed, Fashahah, Lagu,* and *Suara*. The *Tajwed* dimension has four items: *Makharij al-hurf* (The place where the *hijaiyah* or Arabic letters come out when pronounced), *shifah al-hurf* (The nature or characteristics of the *hijaiyah* letters when pronounced), *ahkam al-hurf* (The law of pronouncing *hijaiyah* letters), and *ahkam al-mad wa al-qashr* (The law of lengthening or shortening recitations in the Quran). The *Fashahah* dimension comprises five items: *ahkam al-waqf wa al-ibtida'* (The law of stopping and starting in reciting the Quran so that the meaning remains intact), *mura'ah al-hurf aw al-harakah* (Paying attention to the letters and *harakat* or punctuation, when reading the Quran so that the pronunciation is correct), *mura'ah al-kalimah* (Paying attention to every word in the Quranic text to ensure it is not missed or misread), *mura'ah al-ayah* (Paying attention to the entire verse to ensure it is not missed or misread), and *tamam al-waqt* (Allowing enough time to read the Quran, not in a hurry, so that the reading is clear and *tartil*). The *Lagu* dimension consists of five items: an assessment for the first and ending songs; the number and/or composition of the song (specifically for the final round); the transition, complete form, and tempo of the song; rhythm, style and appreciation; and variation. The *Suara* dimension consists of five items: vocals and voice integrity; voice clarity; smoothness/softness; loudness; and breathing regulation. Tables 3.1 to Table 3.4 are assessment rubrics for each dimension in English version.

Regarding the scoring system for each item, the criteria for participants' errors in the recitation of the Qur'an were referenced when establishing the rules of the MTQ competition. Within the dimensions of *Tajwed* and *Fashahah*, two categories of errors are identified. The former, termed "*khafi*," is considered a minor mistake, while the term "*jali*" is used to denote a major mistake. The total maximum score for these two dimensions is 30 points. Participants who make a *jali* mistake in these dimensions will have their score reduced by two points for each mistake, with the same reduction applied for each subsequent mistake. Similarly, in the case of a *khafi* error, the score is reduced by ½ point. It is noteworthy that each incorrect reading is counted as an error, even if the reading is repeated correctly. In the case of *Fashahah* dimension, item *muro'ah al-kalimah*, specifically the addition or omission of a word, it is classified as a three-times *jali* error. Mistakes in item *mura'ah al-ayat,* such as the addition, subtraction, or omission of verses

or more than two words, are classified as five times *jali* error. If the end time marker light remains extinguished and the participant has or has not completed the reading, a deduction of one point will be made for item number five in the *Fashahah* dimension.

Table 3. 3 *The rubric of Tajwed dimension*

No	Item rated	Jali mistake		Khafi mistake		The amount of Jali + khafi mistake	Final score	Notes
		Freq.	Sum	Freq.	Sum			
1	Makharij al-Huruf			… x 1/2				
2	Shifah al-Huruf			....x 1/2				
3	Ahkam al-Huruf			… x 1/2				
4	Ahkam al Mad wa al-Qashr	....x 2		… x 1/2				
Maximum score. 30				Final score = 30 - ..... = ...				

In this dimension, there is one item that has two types of mistakes. Item number four, *Ahkam al-Mad wa al-Qashr*, has *Jali* and *khafi* error types. Accordingly, the number of points deducted will be greater if the student makes a *jali* error. Likewise, if the student makes a *khafi* error, the point deduction will be the same as the other items. Therefore, every single *jali* error made in item number four is considered to have made four errors in this dimension. This is based on the rubric that, under normal conditions, all errors will result in a ½ point deduction. While the *jali* error causes a deduction of two points.

Table 3. 4 *The rubric of the Fashahah dimension*

No.	Item rated	Jali mistake		Khafi mistake		The amount of Jali + khafi mistake	Final score	Notes
		Freq.	Sum	Freq.	Sum			

1	Ahkam al-Waqf or al-Ibtida	…. x 2	… x ½
2	Mura'ah al-Huruf wa al-Harakat	….x 2	
3	Mura'ah al-Kalimah	6	
4	Mura'ah al-Ayat	10	
5	Tamam al-Waqt*		
Maximum score. 30		.....	Final score = 30 - ..... =

*) The maximum deduction of 1 point

In this dimension, there is one item that has two types of mistakes. The item is item number one, called *Ahkam al-Waqf wa al-Ibtida*, which has a *jali* and *khafi* type of error. As mentioned in the *Tajwed* dimension, the deduction for every error in these two types of errors is different. In the normal condition for items one and two, every single error is considered to have a 2-point deduction. However, specifically for item number one, if the performer makes a *khafi* error, the point deduction will only be a ½. Therefore, in item number one, if the performer makes a *khafi* error, the number of errors is considered as ¼ normal mistake in this dimension.

Table 3. 5 *The rubric of Lagu dimension*

No	Item rated	Score		Score reduction	Total	Procurement	Note
		Max	Min				
1	First and ending song	5	½				
2	Number of songs/compositions	5	½				
3	The transition, complete form, and tempo of the song.	5	½				

4	Rhythm, style, and appreciation	5	½
5	Variations	5	½
Maximum score. 25		Final score = 25 - ..... = ...	

Table 3. 6 *The rubric of Suara dimension*

No	Item rated	Score		Value reduction	Sum	Procurement	Note
		Mak	Min				
1	Vocals and voice integrity	3	1				
2	Clarity or clarity	3	1				
3	Smoothness and softness	3	1				
4	Loudness	3	1				
5	Breathing regulation	3	1				
Maximum score. 15		Final score = 15 - ..... = ...					

Within the *Lagu* dimension, errors affecting the score occur when the number of songs is less than the minimum limit or incomplete, when the song composition is inappropriate, when the song transitions are incongruous, the rhythms, styles and variations of songs are not harmonious, and when the tempos are inconsistent. In the *Suara* dimension, the errors in this case are rough, hoarse, incapable of high, discordant, groaning, inconsistent, and uncontrollable breathing during performance. Within the *Lagu* dimension, the maximum total score is 25 points, while the minimum is 2.5 points. In the *Suara* dimension, 15 points is the maximum total score, and 5 points is the minimum score. Scoring is done by deducting ½ point for each error. Unless the number of songs is less than the minimum limit, 2 points are deducted. Additionally, if participants incorporate verses after the cessation of the designated time interval, a deduction of one point is made from the initial song's total score. The range of scores for each item is confined to a maximum of five points and a minimum score is one and half point. In the *Suara* dimension,

the range of scores for each item is from 1 to 3. A score of 1 represents the lowest possible score, while a score of 3 representing the highest possible score.

Due to the varying weights of the items across dimensions, it is imperative to standardize the instrument's data processing. A likert rating scale formula was created for each item using the Quranic recitation assessment data from this study. Not only that, the results from the MTQ competition at the national level, which data is accessible on the website "www.musabaqah.id.", were used as the supporting reference and comparison for the Likert rating scale formulation.

Table 3. 7 *The Likert scale formula across all dimensions*

Tajwed	**Items**	**1-4**				
	Likert rating scale	1	2	3	4	5
	The number of errors	≥ 7	5-6	3-4	1-2	0
Fashahah	**Items**	**1-2**				
	Likert rating scale	1	2	3	4	5
	The reduction of errors	≥ 4	3 - < 4	2 - < 3	1 - < 2	0 - < 1
	Items	**3-5**				
	Likert rating scale	1		2		3
	The number of errors	≥ 2		1		0
Lagu	**Items**	**1-5**				
	Likert rating scale	1	2	3	4	5
	The reduction of score	> 3	> 2-3	> 1-2	> 0-1	0
Suara	**Items**	**1-5**				
	Likert rating scale	1		2		3
	The reduction of score	> 1		> 0-1		0

3.7. Data Analysis

In this study, the obtained data were analysed using the Many-Facet Rasch Measurement (MFRM) model. The initial step was the collection and documentation of the assessment rubric that had been completed by raters. The data in the form of scores from the assessment rubric were transformed according to the likert rating scale formula (see table 3.7) to be collected as raw data in Microsoft Excel 2019 version. Demographic data for each participant was entered into Microsoft Excel and labelled. Any missing or incomplete data were recorded, adjusted, and/or excluded from the analysis.

The second step was formulating the coding. Since the data analysis utilized MFRM, Facet software, which was developed by Linacre, was used. The researcher formulated a coding formulation containing a programming command language that explains the data conditions and the analysis commands needed so that the data can be analysed with the software. Data previously computed in Microsoft Excel was also included in the coding formulation. Once the coding was complete, it was then entered into the MFRM model for analysis using Facet software 3.66.0 version (Linacre, 2023).

However, it is important to note that the data obtained in each dimension has a different meaning and type of Likert rating scale. These differences require the data to be analysed separately. Not only that, several items in the *Fashahah* dimension have different score weights compared to other items. This is different from other dimensions where all items have the same score weight. According to Engelhard (2013), Masters (2016), and Andrich (2019), the case in the *Fashahah* dimension indicates that the data is classified as a compound ranking type, which requires a specific model called the partial credit model (Engelhard, 2013; Masters, 2016; Andrich, 2019). Consequently, the coding formulation in the *Fashahah* dimension differed from that of other dimensions, where they used a model called the rating scale model (Andrich, 2019) but can still be accommodated with MFRM in the Facet software. The only difference was in how the item's thresholds were estimated. For the *Fashahah* dimension, the estimation was done separately for each item. For the other dimensions, one set of thresholds was used for all items (Andrich, 2019). All this analysis referred to the Many-Facet Rasch Measurement formula as it stated by the explanation of Engelhard (2013).

Many-Facet Rating Scale Model

$$\ln\left(\frac{Pnmik}{Pnmik-1}\right) = \theta n - \lambda m - (\delta i + \lambda k) = \theta n - \lambda m - \delta i - \lambda k$$

Many-Facet Partial Credit Model

$$\ln\left(\frac{Pnmik}{Pnmik-1}\right) = \theta n - \lambda m - \delta ik$$

Definition:

$Pnmik$: Probability of a specific rating being assigned

θn: Student's (performer) Quranic recitation ability n

λm: Scoring severity of raters (judges) m

δi: Item's difficulty

λk: Rating category's difficulty

δik: Difficulty of step k of item i (assuming unique scale structure of each item)

Following the formula used, the next step was interpreting the results output from the Facet software. Since this study will focus on the three areas, which are validity, reliability, and fairness, the specific indicators will be sequentially referred to the source of evidence in each cluster chosen by this study. Before that, the data landscape conditions will be detailed by considering the summary statistics report. For validity, the analysis was intended to assess the internal structure and response process of the rubric. In the internal structure evidence, the model data-fit analysis of the items followed the rules for acceptable values of "infit and outfit Mean Square (MNSQ), Z standard (ZSTD), and Point Measure correlation statistics" set by some researchers. According to Boone et al. (2014) and Engelhard and Wind (2018), the expected value for infit and outfit mean square is between 0.5 to 1.5 (Boone et al., 2014; Engelhard & Wind, 2018). A value close to 1 indicates a good fit between the item/person; otherwise, they are considered a misfit and require further investigation (see Table 3.8). The expected value for infit and outfit ZSTD is between -2.0 to +2.0, while the point measure correlation should be in the range of positive values (Boone et al., 2014). The detailed fit category and its interpretation were referred to Engelhard & Wind (2018) for mean square interpretation and Linacre (2002) for the z-

standard interpretation. Following the analysis steps outlined by Wang et al. (2022), the analysis will also measure each facet by analysing a Wright map, particularly in explaining the order of the items and their impact on the level of performers' ability and rater severity level in the assessment (Wang et al., 2022). In examining the unidimensionality of the rubric, a minimum raw variance value of 20% set by Reckase (1979) will be used as the standard in this study (Reckase, 1979).

In the response process evidence, the category scale statistics report is used to ensure the order of the category is increasing monotonically from the lowest to the highest. The Rasch Andrich threshold is used to assess whether the distance between adjacent thresholds is acceptable, as suggested by Linacre (2002). The counts used, average measure, the outfit mean square, and the report from the probability curves will be considered to ensure the response process given to the rubric scale category in this study.

Table 3. 8 *Fit category based on mean square error (MSE)*

MSE	Interpretation	Fit category
$0.50 \leq MSE < 1.50$	"Productive for measurement."	A
$MSE < 0.50$	"Less productive for measurement but not distorting of measures."	B
$1.50 \leq MSE < 2.00$	"Unproductive for measurement but not distorting of measures."	C
$2.00 \leq MSE$	"Unproductive for measurement and distorting of measures."	D

Engelhard and Wind (2018) (cited with permission)

Table 3. 9 *Guideline for the interpretation of Z Standard (ZSTD)*

Standardized value	Interpretation
≥ 3.0	"Data very unexpected if they fit the model (perfectly), so they probably do not. But, with large sample size, substantive misfit may be small."
2.0 – 2.9	"Data noticeably unpredictable."
-1.9 – 1.9	"Data have reasonable predictability."

≤ -2.0	"Data are too predictable. Other "dimensions" may be constraining the response patterns."

Linacre (2002) (cited with permission)

The next analysis is reliability. The reliability evidence will rely on the concept that the reliability index is the ratio of true score variance to observed variance based on the latent measures (Linacre, 2023). The Reliability $= \frac{SD^2 - MSE}{SD^2}$, where SD is the standard deviation of Rasch measures and MSE is the average of squared standard errors of Rasch measures (Engelhard & Wang, 2021; 2024). This index reflects how different the measures of each facet are, and its value ranges from 0 to 1. High values (close to one) indicate a good representation of Rasch measures across the entire range of the latent scale. Referring to Fisher (2007), the accepted reliability index is a value greater than 0.65 (Fisher, 2007). The separation index is determined to assess the distribution of respondents based on their ability levels in the measured variables (Bond et al., 2021). The separation index, if greater than 2, indicates a good distribution in the data obtained (Bond et al., 2021). To check the precision of the assessment, the standard error indices are used. The low index reported shows that a low measurement error has occurred (Linacre, 2023). The inter-rater agreement from the observed and expected agreement percentage will be considered to check the consistency between raters in the assessment (Linacre, 2023).

To assess the fairness of the assessment, the evidence collected will start with the analysis of model-data fit for persons by following the indicator consideration as mentioned in the validity section. The goals are to identify any systematic biases that could result in unexpected responses. The Differential Item Functioning (DIF), Differential Rater Functioning (DFR), or interaction that can cause systematic bias and inaccurate measures will be pointed. This is important because it can threaten validity, reliability, generalizability, fairness, and ethical issues (Wang et al., 2022). Therefore, calibration is required to put both groups on a common scale before performing bias analysis. FACETS uses a two-stage process: calibrate all facets onto the same scale and estimate the size of the bias/interaction/ DIF/DRF effect (Linacre, 2023). All this evidence is collected to accommodate the irrelevant barriers that occurred in the assessment. Therefore, this study can provide a meaningful report, finding, or suggestion to improve the quality of the rubric.

Besides the psychometric statistics report, the notes from raters regarding the participant performances as qualitative data were used to support the findings from the MFRM analysis. For instance, in the validity analysis, the MFRM analysis reported that

some items are indicated to have misfit or overfit indices. Then, the qualitative data will be used to understand the intended aspect assessed by the items by considering the assessment pattern given by the rater. In analysing the rater bias or unexpected responses given by the raters, the qualitative data used provides concrete evidence that represents the unfair responses that occurred in the assessment. Hence, by having the qualitative reports in this assessment were intended to be used to enrich the data analysis while providing a deep analysis of the assessment. By the end, the answer to all research questions was intended to conclude whether the quality of the Quranic recitation rubric assessment is valid, reliable, and fair to potential assessment bias.

3.8. Ethical Considerations

The ethical considerations of data collection were upheld in this study, following academic standards. Regarding the consent given, all data collected were managed following established research ethics standards as suggested by Strike (2006), Milner (2007), and Holmes (2014). The use of the *Pesantren* as a research location and the involvement of participants in the study have received official informed consent. Research recommendations from the National LPTQ and the Banten Province LPTQ ensure that this research has been considered and is feasible to be conducted. Regarding anonymity, names and any information beyond that required for the data have been anonymized to maintain privacy and confidentiality. All the ethical provident can be seen in the appendices.

CHAPTER IV
RESEARCH RESULT AND DISCUSSION

This chapter presents a comprehensive and detailed analysis of the research findings derived from the collected data. The Quranic recitation performance of 50 students was assessed by 16 judges using a Quranic recitation assessment rubric with four dimensions and 19 items. This produced 3.760 quantitative responses and 674 qualitative responses. The quantitative data responses were analysed using FACETs software, which is specialised application that uses Many-Facet Rasch Measurement (MFRM). The qualitative responses supported the quantitative results, helping to answer the research questions. To systematically analyse the data and achieve the research objectives, this chapter is divided into three parts. First, the overview of the MFRM analysis results provides a general description of the MFRM analysis output, which contains four dimensions in the Quranic recitation assessment rubric: *Tajweed, Fashahah, Lagu*, and *Suara*. Second, the findings present the answers to the research questions posed by this study. Third, the discussion will integrate and interpret these findings in light of the answers to research questions. In this chapter, to use terminology consistent with MFRM analysis from Facet software, the students as performers will be called ratees, and the judges will be called raters.

4.1. Overview of Many Facet Rasch Measurement (MFRM) Results

A general report of the results of the analysis that was conducted with MFRM is presented in this subsection. Before carrying out an in-depth analysis and substantive interpretation to answer the research questions, information about the overall data landscape is provided to develop an initial understanding. The discussion will sequentially cover the data landscape, summary statistics across dimensions, and the distribution of participants' demographic information.

4.1.1. Data Landscape

The Quranic recitation assessment rubric consists of four dimensions: *Tajwed, Fashahah, Lagu,* and *Suara*. These dimensions are aspects of assessment to measure the quality of a performer's Quranic recitation. The resulting assessment dataset structure differs because the different conditions of each dimension require separate analysis. Therefore, the data landscape is presented separately.

The data in the *Tajwed* dimension was calculated based on the responses of Quranic

recitation experts (n = 4) who assessed 50 participants. The four items were measured, including *makharij al-hurf, shifah al-hurf, ahkam al-hurf*, and *ahkam al-mad wa al-qashr* in the range of likert rating scale from 1 to 5. A total of 800 quantitative responses were produced, and 194 qualitative responses were also included.

In the *Fashahah* dimension, data were obtained from the Quranic recitation performance of 50 participants, who were assessed by four raters. The assessment focused on five items: *ahkam al-waqf wa al-ibtida', mura'ah al-hurf aw al-harakah, mura'ah al-kalimah, mura'ah al-ayah,* and *tamam al-waqt*. From an expected 1.000 quantitative responses, 40 were missing, resulting in 960 responses. Additionally, 124 qualitative responses were obtained from the notes provided by raters. The results of the analysis in this dimension used a different model, namely MFRM, which has the partial credit model, because the items had a different likert rating scale as suggested by Engelhard (2013), Masters (2016), and Andrich (2019). Rating scale 1 to 5 for items F1 and F2, rating scale 1 to 3 for items F3, F4, and F5.

In the *Lagu* dimension, data were obtained from 50 participants who were assessed by four judges based on five items of criteria. These items were assessed using a scale of 1 to 5 and consisted of the following: an assessment for the first and ending songs; the number and/or composition of the song (specifically for the final round); the transition, complete form, and tempo of the song; rhythm, style, and appreciation; and variation. A total of 200 data sets with 1000 quantitative responses and 179 qualitative notes were provided.

The last dimension measured in the Quranic recitation performance was the *Suara* dimension. In this dimension, data were also obtained from four raters who assessed 50 participants on five assessment items. These items included vocals and voice integrity; voice clarity; smoothness/softness; loudness; and breathing regulation. The assessed process from the measurements carried out on a scale of 1 to 3 produced 1.000 quantitative responses and 177 qualitative responses.

Overall, the Facet software processed a total of 3.760 quantitative data responses. Of the total data that should have been able to reach o3.800 responses, 40 were missing, all of which occurred in the *Fashahah* dimension. According to the data compiled in Microsoft Excel, two raters did not provide an assessment for eight participants. Nevertheless, data reports in the form of qualitative responses were given to seven of the eight participants who did not receive quantitative assessments. Therefore, although MFRM can still accommodate this issue, 674 qualitative responses provided by raters will be used as supporting data to enrich the data interpretation.

4.1.2. Statistical Summary of the Datasets Across Dimensions

Tajwed dimension

Table 4. 1. *Tajwed measurable data summary*

Category	Score	Expected score	Residual		Standardized residual	
			Mean	S.D.	Mean	S.D.
3.72	3.72	3.72	0.00	0.89	0.00	0.99

Responses used for estimation	Count	Mean	S.D.	Params
	800	3.72	1.07	59

Count of measurable response	800
Data log-likelihood chi-square	1922.1549
Approximate degrees of freedom (d.f.)	741
Chi-square significance probability (*p*)	0.0000
Raw-score variance of observation	1.14 (100%)
Variance explained by Rasch measurement	0.35 (30.56%)
Variance of residuals	0.79 (69.44%)

Table 4.1. is a measurable data summary from Facets software that provides detailed information related to the assessment given. The output from the analysis of MFRM, it can be known there were 800 responses used for estimation, with the mean of 3.72 and standard deviation of 0.89. The residual value has a mean of 0.00 and a standard deviation of 0.89. The standardized residual has a mean of 0.00 and a standard deviation of 0.99. According to Linacre (2023), residual is the difference between the actual score and the expected score, while the standardized residual is the residual value that has been divided by the standard deviation which functions to see the suitability of the data to the model (Linacre, 2023). These results show that the MFRM model used in estimation for 800 responses is quite good at fitting the data. The reason is that the residual value of this data is close to zero, while the residual standard deviation value is within reasonably acceptable ranges (Linacre, 2023). Therefore, these results indicate that there is no indication of extreme data variation between the observed data and the data predicted by the model.

According to Engelhard and Wind (2018), log-likelihood chi-square is one of the outputs that explains the global fit data measure to show how well the data fits the model. Degrees of freedom (d.f.), which is the amount of freedom in the model related to the number of items and respondents, is also a component that needs to be reported together

with the log-likelihood chi-square (Engelhard & Wang, 2024). From Facets software, the chi-square log-likelihood obtained on the *Tajwed* dimension is at 1922.1549 with a degree of freedom (d.f.) at 741. This produces a significance value (*p*) probability of 0.0000, indicating that there is a significant difference between raters, items, and ratees in the measured model (Engelhard & Wang, 2024; Eckes, 2015). These results confirm that the diversity of data, such as differences between raters or variations between ratees, captured by the model is not the result of chance, but reflects the existence of real facets.

Apart from that, it should be noted that the variance explained by Rasch measurement value obtained was at 30.56%, while the unexplained variance or residue was at 69.44%. Referring to the existing standard guidelines, experts have different opinions regarding the minimum standard of the value of variance explained by the Rasch measure. Reckase (1979) set 20% as the minimum standard and others consider that 40% or 50% (Bond & Fox, 2015; Fisher, 2007) is the minimum standard that must be met to show that the data is productive, measure unidimensionality, and has a reasonable prediction (Reckase, 1979; Wind, 2023). In this context, this study followed the standard of Reckase (1979), which states that the value generated from the variance explained by the Rasch measure is acceptable. Although it must be noted that the Rasch model still accounts for more than half of the data variation (69.44%), which cannot be explained by the model and needs to be considered, as suggested by other experts. The reason why the Reckase (1979) standard is being used because the cause is suspected to be due to other factors such as the inconsistency of the rater in giving scores, the quality of the items, the abilities' pattern of the ratees, or the use of suboptimal scale categories to measure responses (Reckase, 1979; Linacre, 2023). Thus, although statistically the data on the *Tajwed* dimension have a good basic level of fit with the MFRM, an in-depth analysis of rater performance, item function, scale structure, and ratee performance needs to be conducted to provide a more in-depth analysis.

Fashahah dimension

Table 4. 2. *Fashahah measurable data summary*

Category	**Score**	**Expected score**	**Residual**		**Standardized residual**	
			Mean	**S.D.**	**Mean**	**S.D.**
3.59	3.59	3.59	0.00	0.67	0.02	0.85
Responses used for estimation				**Count**	**Mean**	**S.D.** **Params**

	720	3.59	1.05	63
Responses with invalid elements	192	3.00	0.00	0
Responses in one extreme score	48	4.00	1.00	3
Count of measurable response	960	3.49	0.97	66
Data log-likelihood chi-square	857.8340			
Approximate degrees of freedom (d.f.)	657			
Chi-square significance probability (*p*)	0.0000			
Raw-score variance of observation	1.10 (100%)			
Variance explained by Rasch measurement	0.64 (58.67%)			
Variance of residuals	0.45 (41.33%)			

Table 4.2. is a measurable data summary that explains the quality and suitability of the data regarding rater responses to Rasch model. The mean category, score, and expected score values are at 3.59, indicating a symmetrical and centered data distribution. The residual value of 0.00 and the standardized residual of 0.02 show that there is no significant difference between the observed and predicted values by the MFRM analysis. Referring to Linacre (2023), this indicates that the model successfully estimated the data.

Further analysis of the data quality can be seen through the chi-square log-likelihood value which is at 857.8340 with a degree of freedom of 657, producing a significance value (*p*) of 0.00. Statistically, these numbers explain that there is a significant difference between the model and the data. However, it should be noted that in the context of MFRM, these results do not necessarily mean that the data is poor. According to Linacre (2023), when processing large amounts of data, the chi-square value is often statistically significant. But in practice, it does not always indicate a major problem if the residual variance remains within reasonable limits (Linacre, 2023).

Highlighting the results of variance explained by Rasch measurement and variance of residuals, 58.67% of the data can be explained by the Rasch model, while 41.33% of the data cannot. These results suggest that the amount of data variation that can be explained is quite feasible in the context of human assessment, as it exceeds the 20% standard (Reckase, 1979; Wind, 2023). However, the percentage of residual variance needs to be considered because it indicates the presence of inter-rater bias or ambiguity in the rubric that the model did not reveal (Engelhard & Wang, 2024).

This study yielded some interesting findings related to the data. Of the total 960 data

points recorded, 48 were extreme responses and 192 were invalid responses. The extreme responses are likely due to participants who consistently receive the highest or lowest scores from the rater. Another possible cause is the tendency toward bias in certain aspects or the appearance of participants who are indeed very good or very poor in his or her performance. The 192 invalid responses indicate the need for further investigation, as this number is quite large-20% of the total responses. These invalid responses certainly have implications for the overall quality of the measurement. Thus, the handling of these responses needs to be investigated further so as not to affect the parameter estimates in the Rasch model.

Lagu dimension

Table 4. 3 *Lagu measurable data summary*

Category	**Score**	**Expected Score**	**Residual**		**Standardized Residual**	
			Mean	**S.D.**	**Mean**	**S.D.**
4.31	4.31	4.31	0.00	0.62	0.04	0.94

Responses used for estimation	**Count**	**Mean**	**S.D.**	**Params**
	1000	4.31	0.97	60

Data log-likelihood chi-square	1434.7861
Approximate degrees of freedom (d.f.)	940
Chi-square significance probability (*p*)	0.000
Raw-score variance of observation	0.95 (100%)
Variance explained by Rasch measurement	0.56 (58.94%)
Variance of residuals	0.39 (41.06%)

The results of the data analysis obtained for the *Lagu* dimension were analysed using MFRM and reported in table 4.3. This table is a measurable data summary which fully explains that the data structure used is quite good, with a response distribution that can be interpreted statistically. A total of 1.000 responses were used for estimation, the processing results provide a broad picture of the response patterns in the dataset. The mean value of 4.31 explains that high scores are often given to the measured items. While the standard deviation value is 0.97, which according to Fisher (2007), is excellent at providing and capturing response variability in the *Lagu* dimension.

From the perspective of Chi-Square log-likelihood, the value of 1434.78 with 940

degrees of freedom and a *p* value of 0.000 indicates a significant difference in response distribution, confirming that the response pattern reflects real differences (Linacre, 2023). In terms of variance, 58.94% is explained by Rasch measures, while 41.06% which is the residual variance indicates that other factors may influence the response given. Referring to the Reckase standard (1979), this value meets the standard. Moreover, according to the explanation of Bond and Fox (2015), the variance explained figure of more than 40% is considered good in indicating that the responses have measurement stability. Additionally, the number of parameters used in the estimation is 60. This indicates that the model has a complex structure for adjusting the measurements to the obtained data (Linacre, 2023). Although simplifications can be made to improve estimation efficiency, the number of parameters remains proportional to the number of responses used.

Suara dimension

Table 4. 4 *Suara measurable data summary*

Category	**Score**	**Expected score**	**Residual**		**Standardized residual**	
			Mean	**S.D.**	**Mean**	**S.D.**
2.40	2.40	2.40	0.00	0.56	0.01	0.99

Responses used for estimation	**Count**	**Mean**	**S.D.**	**Params**
	980	2.40	0.74	57

Data log-likelihood chi-square	1430.4139
Approximate degrees of freedom (d.f.)	923
Chi-square significance probability (*p*)	0.0000
Raw-score variance of observation	0.54 (100%)
Variance explained by Rasch measurement	0.22 (41.30%)
Variance of residuals	0.32 (58.70%)

In general, the results of the quantitative dataset measured by the MFRM can be estimated quite well. This statement can be seen in the Facet's report shown in Table 4.4. The average score obtained with the average score expected by the model is 2.40, which indicates a good match between the data obtained and the model (Linacre, 2023). The standard deviation is reported on the 0.74 indices. The residual and standardized residual values have averages and standard deviations that are close to ideal, indicating that the data is fit with the model and can provide estimation of data with high accuracy (Linacre, 2023).

Regarding the sample size used for response estimation, 980 responses and 57 parameters in the model allow parameter estimation with a high level of precision. Chi-square analysis of the log-likelihood data yielded a value of 1430.4139 with a degree of freedom of 923, and a significance probability of p = 0.000. Referring to Linacre's explanation (2023), these results explain that there is a significant difference between the observed data and the model used.

From a variance perspective, there is a 0.54 variation in raw scores representing 100% of the total data variability. The model can explain up to 0.22 or 41.30% of the variance, while up to 0.32, or 58.70%, of the residual variance cannot be explained by the model. According to Reckase (1979) and Wind (2023), this percentage is quite good because the value exceeds the ideal standard of 20%. These results indicate that the model can sufficiently explain the variability of the data, some aspects remain unexplained. Therefore, the data in the *Suara* dimension is considered to have a fairly good level of fit with the model, as indicated by the significance of the chi-square probability and the high percentage of residual variance, which can be considered for further evaluation.

Overall, the dataset obtained in each dimension was processed well by MFRM through Facet software. The statistical results for several indicators in each dimension have shown good values, with many close to the ideal standard. However, it must be acknowledged that several indicators, such as the percentage of variance explained by measures, can be improved. Therefore, with the favourable results from the good dataset conditions, further analysis can be carried out to provide valuable insights, especially for evaluating the quality of the Quranic recitation assessment rubric.

4.1.3. Statistical Participants' Distribution across their Demographic Information

In this study, some demographic information was collected from participants (rater and ratee). For the rater group, demographic information included gender, provincial origin, age, length of stay at the *Pesantren*, position at the *Pesantren*, experience, and field of expertise in the MTQ competition. For the ratees, information includes gender, province of origin, age, grade level, length of study at the *Pesantren*, and frequency of participation in the MTQ competition. The diversity of the distribution of each demographic group will be presented in this section, referring to the results of the MFRM analysis.

Distribution grouping in statistical analysis is generally done by referring to the mean and standard deviation. Standard deviation is a measure of data dispersion, indicating the extent to which values in a data set are spread from the mean (Wright & Stone, 1979). In

the context of this study, each facet, particularly raters and ratees, has its mean and standard deviation. Therefore, the distribution of demographic backgrounds for all participants can be categorized by referring to these results. However, in Rasch analysis, group division can also be done by referring to the estimated value of the separation index in each summary report. Sumintono & Widhiarso (2015) provide the following equation to view the grouping more clearly: $H = \frac{[(4 \text{ x Separation}) + 1]}{3}$. The results for these two methods are different. Grouping based on separation values tends to be more precise in assessing data discrimination. Meanwhile, grouping based on the mean and standard deviation values distributes the data evenly according to its distribution. Still, the researcher understands both methods good because they offer a comprehensive overview of how participants are spread out in this study.

Referring to the results of the MFRM analysis, specifically in the outputs of tables 7.1.1 and 7.2.1, the H values produced vary. For the *Tajwed* dimension, the ratee and rater values were 3.06 and 3.82, respectively. The *Fashahah* dimension had values of 1.72 and 3.83 for the ratee and rater, respectively. The *Lagu* dimension produced values of 3.97 and 3.72 for each ratee and rater group, respectively. The *Suara* dimension produced values of 3.10 and 4.75 for the ratee and rater groups, respectively. Recognizing the significant variability in the grouping results suggested by the model, this study selected a grouping method based on the mean and standard deviation. The reason is based on this overview section explains the distribution of demographic information for each individual. Therefore, although grouping using separation values can offer more precise discrimination, low scores on the *Fashahah* dimension, particularly in the ratee group, below 2, are assumed to have a less diverse and provide an inconsistent number of groups distribution of demographic information in each dimension.

Choosing a grouping method based on the mean and standard deviation, Table 4.5 shows the extracted report from each dimension. The resulting groups are four groups with varying levels: very high, high, moderate, and low. The formation of these participant groups is based on specific levels, determined using their Logit Value Person (LVP) (both rater and ratee). These levels are determined by referring to ranges derived from calculating the mean and standard deviation of the LVP as presented in Table 4.5. The distribution will be explained sequentially, starting with the rater distribution across three demographic factors: gender, field of experience related to the MTQ competition, and expertise in the MTQ competition. Furthermore, the discussion will be followed by the ratee distribution focusing on their gender, grade level, years of study, and their frequency of participation

in the MTQ competition. The explanation in this section also considers the four dimensions used in this study.

Table 4. 5 *Item fits statistics across dimensions*

Dimension	Type	Mean	S.D.	The range value of each group based on the Logit Value Person (LVP)			
				Very High	**High**	**Moderate**	**Low**
Tajwed	Rater	-0.65	0.22	LVP > -0.43	-0.43 ≥ LVP > -0.65	-0.65 ≥ LVP > -0.87	LVP ≤ -0.87
	Ratee	0.00	0.68	LVP > +0.68	+0.68 ≥ LVP > 0.00	0.00 ≥ LVP > -0.68	LVP ≤ -0.68
Fashahah	Rater	-2.07	0.33	LVP > -1.74	-1.74 ≥ LVP > -2.07	-2.07 ≥ LVP > -2.40	LVP ≤ -2.40
	Ratee	0.16	0.94	LVP > +1.10	+1.10 ≥ LVP > +0.16	+0.16 ≥ LVP > -0.78	LVP ≤ -0.78
Lagu	Rater	-2.74	0.29	LVP > -2.45	-2.45 ≥ LVP > -2.74	-2.74 ≥ LVP > -3.03	LVP ≤ -3.03
	Ratee	0.00	1.29	LVP > +1.29	+1.29 ≥ LVP > 0.0	0.00 ≥ LVP > -1.29	LVP ≤ -1.29
Suara	Rater	-1.20	0.39	LVP > -0.81	-0.81 ≥ LVP > -1.20	-1.20 ≥ LVP > -1.59	LVP ≤ -1.59
	Ratee	0.08	1.23	LVP > +1.31	+1.31 ≥ LVP > +0.08	+0.08 ≥ LVP > -1.15	LVP ≤ -1.15

Rater

In the context of MFRM analysis, the distribution of raters explains the level of strictness of the raters with which they provide assessments. In a higher-level group, each rater will be considered the rater who has the most severe standards for giving high scores to the ratee's performance. According to Myford and Wolfe (2004), this rater response pattern is considered to be a rater with a high level of severity. Conversely, raters in a lower group are understood to often given high scores to ratees. Myford & Wolfe (2004) refer to these raters as having a low level of severity or being fairly lenient. Referring to Figure 4.1, there is a fairly even distribution of the 16 raters across all dimensions (*Tajwed, Fashahah*,

Lagu, and *Suara*), with dominance occurring in the moderate group at 43.75%. The remaining 56.25% is evenly distributed across the other three groups: very high (18.75%), high (18.75%), and low (18.75%). These results indicate that raters were fairly consistent in their rigor when assigning scores, with most giving moderate ratings. A more in-depth analysis in Figures 4.2, 4.3, 4.4, and 4.5 shows the distribution of each rater across each dimension in more detail, taking into account their respective demographic characteristics.

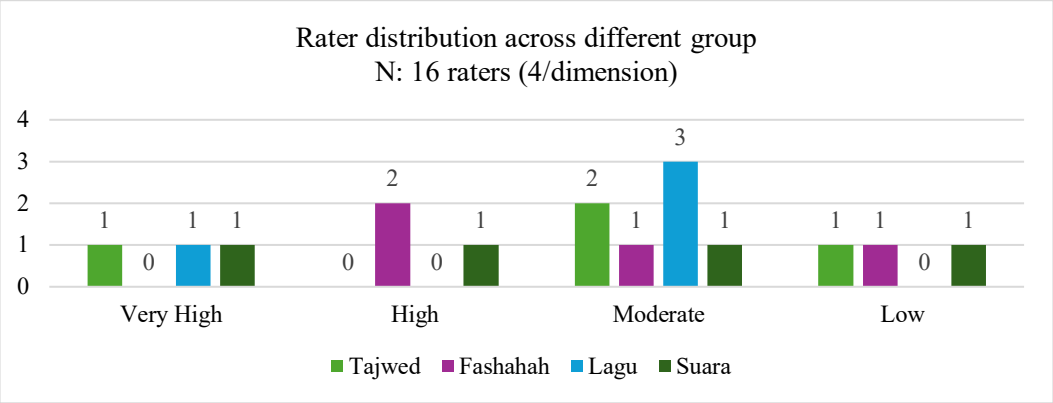


Figure 4. 1 *Rater distribution across different group levels in each dimension*

Tajwed

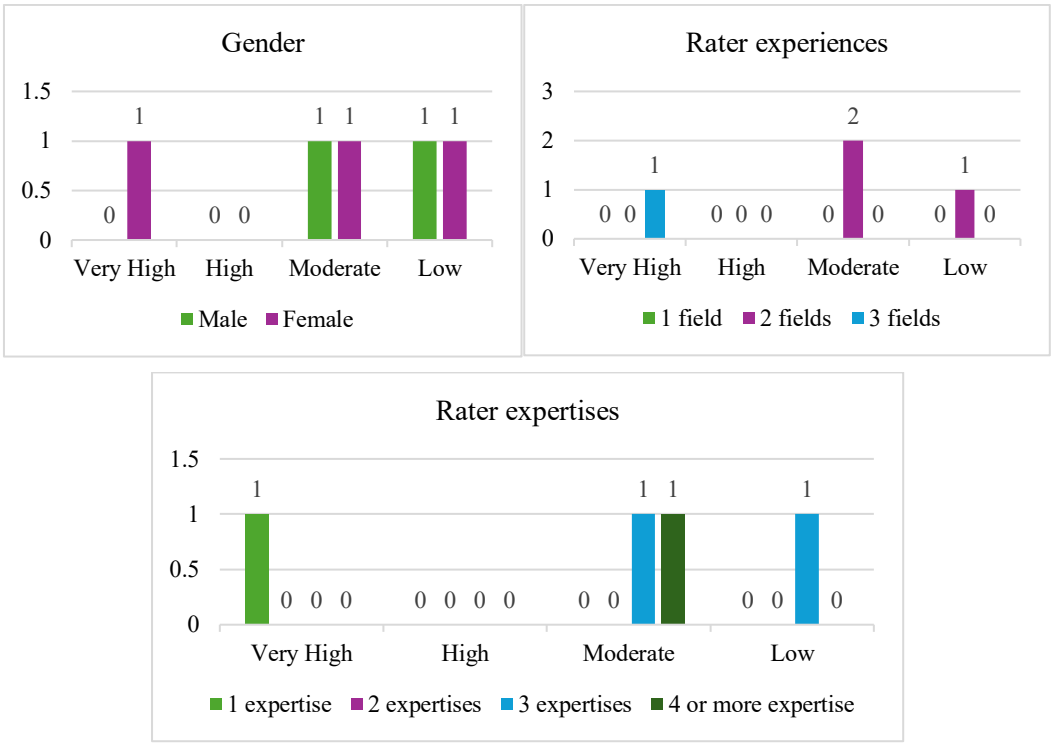


Figure 4. 2 *Rater distribution across different demographics in the Tajwed dimension*

In the *Tajwed* dimension, three demographic information are detailed. These include gender (male and female), rater experience (whether they have experience as a participant, judge, or pedagogue), and rater expertise in the field of MTQ competition, such as *tilawah, tahfidz, tafsir, qiroat, syarhil,* or *murottal*. Overall, the collection of graphs in Figure 4.2 shows that the rater distribution only occurs in the very high, moderate, and low groups, and there are no raters in the high group. In terms of gender, the distribution of male and female raters is evenly distributed across all group levels, with female raters dominating. In terms of the diversity of rater experience in MTQ competitions, raters with at least two experiences, either participant and pedagogue or participant and judge, dominate the rater distribution in the moderate group. Similarly, in terms of the number of raters' expertise, raters with at least three expertise dominate the distribution in the moderate and low groups.

Fashahah

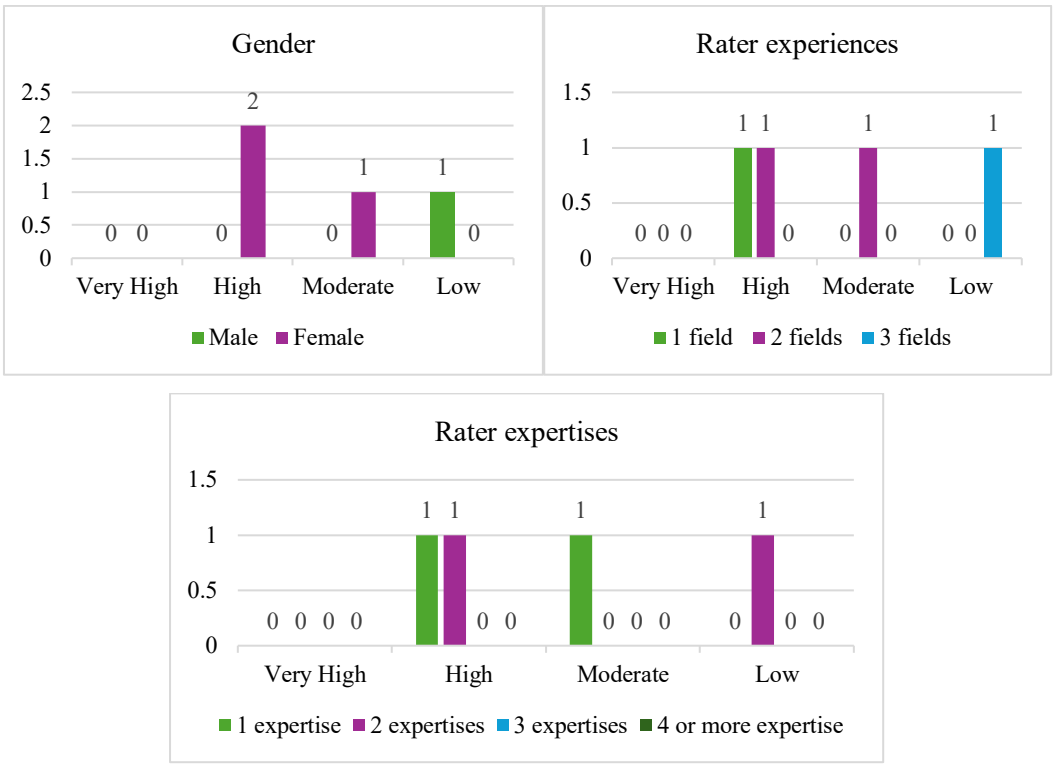


Figure 4. 3 *Rater distribution across different demographics in the Fashahah dimension*

With the same demographic information context, the distribution of raters on the *Fashahah* dimension can be depicted in Figure 4.3. Overall, raters on this dimension are spread across the high, moderate, and low groups. Unlike raters on the *Tajwed* dimension who are not at a high level, none of the raters on the *Fashahah* dimension are in the very

high group. In terms of gender, female raters appear to dominate being in the high group. In terms of rater experience, raters with at least two years of experience dominate being in the high and moderate groups. Meanwhile, raters with one and two years of expertise dominate the distribution on this dimension. The high group is the group that is most occupied by raters when highlighting rater experience and expertise.

Lagu

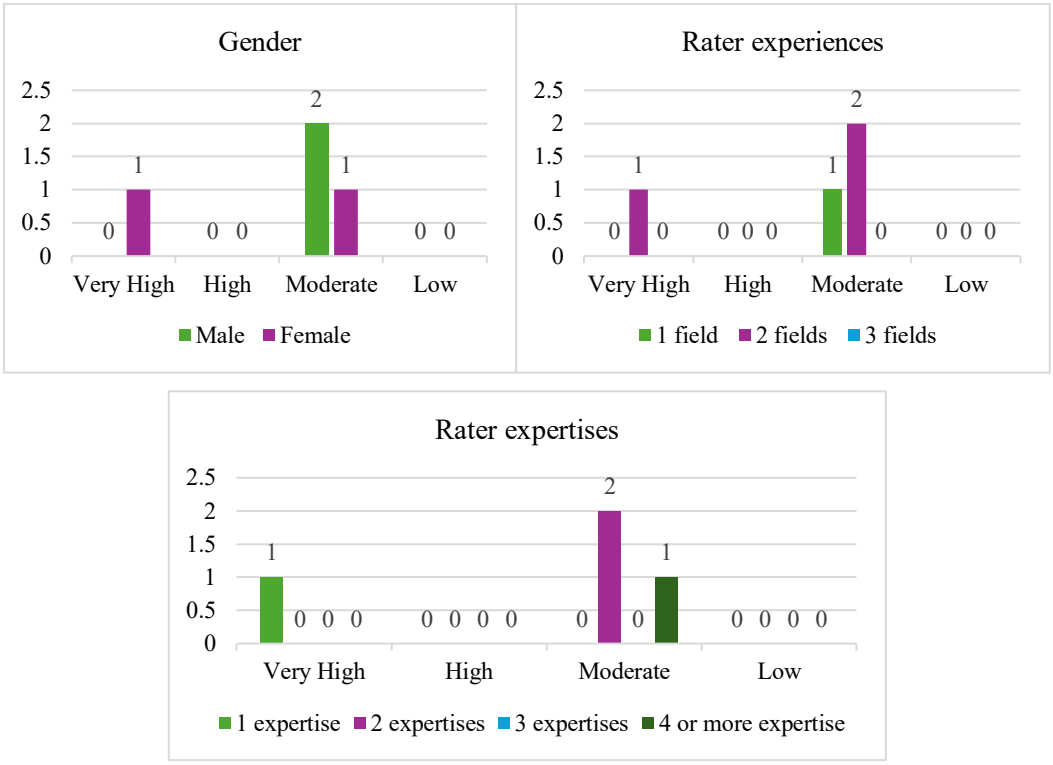


Figure 4. 4 *Rater distribution across different demographics in the Lagu dimension*

In the *Lagu* dimension, the rater distribution can be seen in Figure 4.4, with the distribution occurring only in two groups: the very high and moderate groups. Not a single rater was identified as being in the high or low groups. This indicates that none of the raters were sufficiently lenient in their assessments. On the contrary, the raters were quite severe and moderate in their assessments. In terms of gender, both male and female raters were equally dominant, with the domination of male raters in the moderate group. In terms of experience, the distribution was dominated by raters with two experience in the MTQ. Likewise, in terms of the number of raters' expertise, raters with two expertise dominated this dimension. The distribution of raters that based on their experience and expertise, tends to spread in the moderate group.

Suara

Figure 4. 5 *Rater distribution across different demographics in the Suara dimension*

In the *Suara* dimension, the entire distribution depicted in Figure 4.5 demonstrates the even level of rater rigor across all group levels. From very high, high, moderate, and low levels, at least one rater consistently met all requirements. In terms of gender, female raters appear to dominate the distribution. Similarly, in terms of expertise, raters with two areas of expertise dominate. In terms of experience, raters with one or two areas of experience in MTQ competitions are equally dominant.

Ratee

In the context of MFRM analysis, the distribution of ratees indicates their level of ability in Quranic recitation performance. Each ratee in the high group can be interpreted as having excellent Quranic recitation performance. Meanwhile, ratees in the lower group can be understood as having less good performance in Quranic recitation. The 50 ratees whose distributions were estimated across all dimensions can be visualized in Figure 4.6. Overall, the distribution of ratees across all dimensions was recorded across all group levels. As for the ratees in the *Tajwed, Fashahah*, and *Suara* dimensions, the majority were in the moderate group, and the fewest were in the low group. Meanwhile, for the ratees in

the *Lagu* dimension, the majority were in the high group, with the fewest in the very high group. These results indicate that although the majority of the distribution occurred in the moderate group, considering the number of high and very high groups, the majority of ratees can be assessed as having good ability in Quranic recitation. In more detail, this study also provides information on the demographic characteristics of each ratee across all dimensions shown in Figures 4.7, 4.8, 4.9, and 4.10.

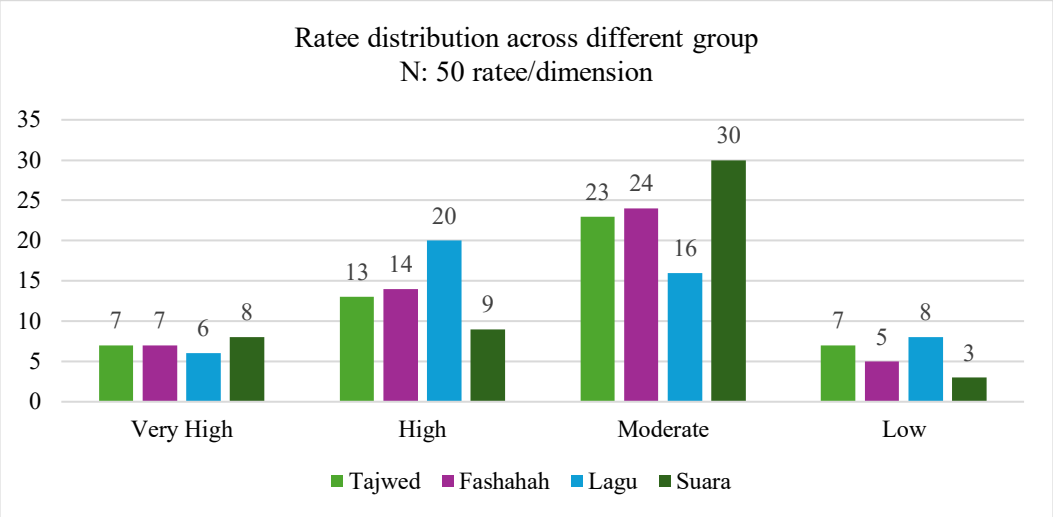


Figure 4. 6 *Ratee distribution across different group levels in each dimension*

Tajwed

In explaining the distribution of ratees at the dimension level, four demographic characteristics were considered. These include gender (male and female), grade level (junior and senior), years of study (1, 2, 3, 4, or 6 years), and frequency of participating in the MTQ competition (1-5 times, 6-10 times, more than 10 times, or never participating). Overall, the collection of graphs in Figure 4.7 shows that the ratees assessed on the *Tajwed* dimension have a highly varied distribution. In terms of gender and grade level, both male and female ratees, both senior and junior, are predominantly spread across the moderate group. In terms of years of study, ratees with a study period of 1 and 3 years are predominantly in the moderate group. Meanwhile, ratees with a study period of 2 and 4 years are in the high group. Ratees with a study period of 6 years are predominantly in the very high group. In terms of frequency of MTQ competition participation, ratees with experience of more than six times tend to be in the high and very high groups. This indicates that their frequency of participation reflects the ratee's excellent Quranic recitation skills. Meanwhile, ratees with less than five MTQ participation experiences, or even those who have never participated, tend to be at a moderate level.

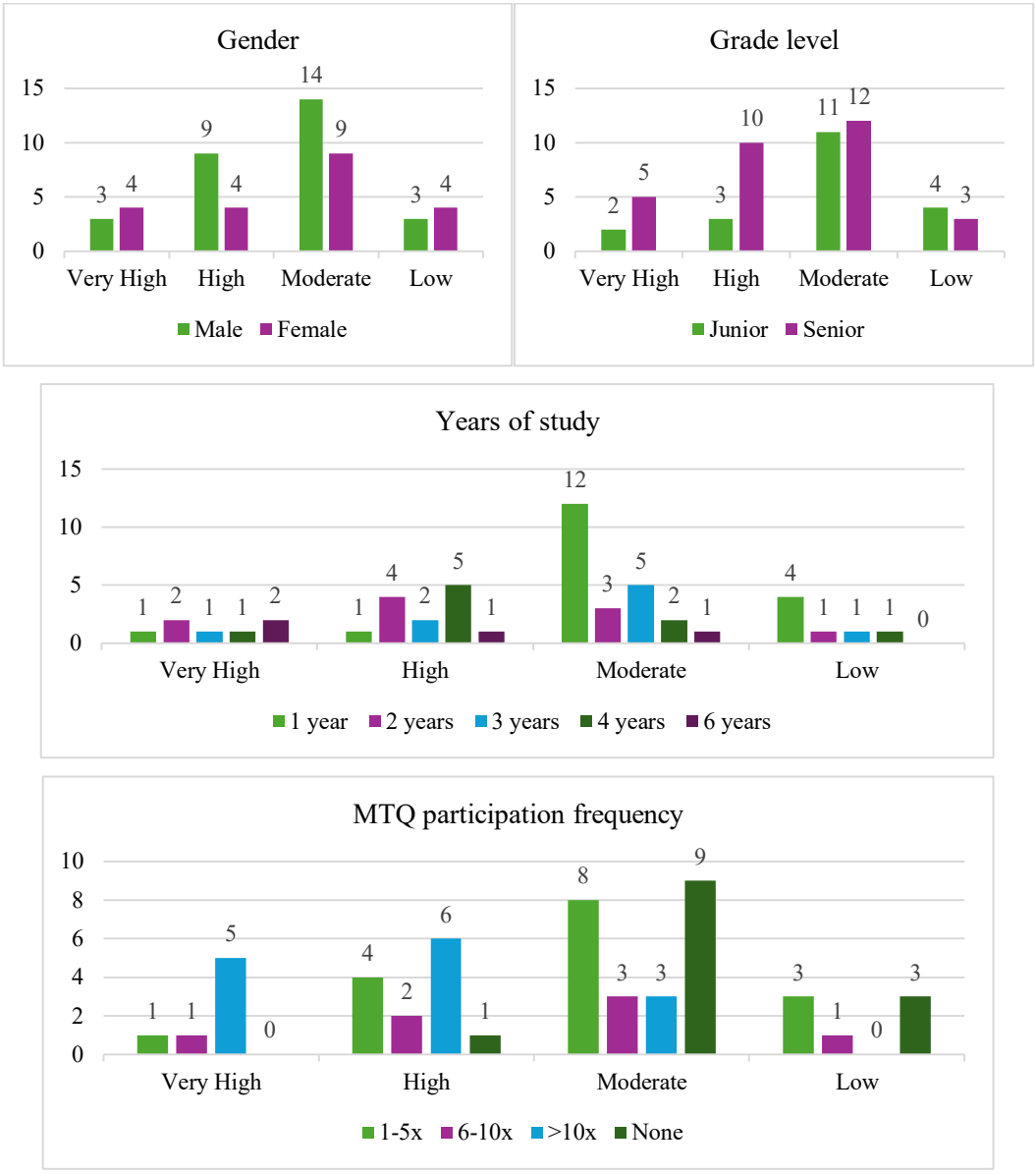


Figure 4. 7 *Ratee distribution across different demographics in the Tajwed dimension*

Fashahah

In line with the analysis of demographic characteristics of ratees in the *Tajwed* dimension, Figure 4.8 also shows a more detailed distribution of ratee groups in the *Fashahah* dimension. Overall, the distribution of ratees appears to be highly varied. In terms of gender, male ratees are evenly distributed across the high and moderate groups, while female ratees are predominantly in the moderate group. In terms of grade level, both junior and senior ratees are in the moderate group. Focusing on years of study, ratees with 1, 3, and 4 years of study are in the moderate group, while raters with 2 years of study are predominantly in the high and moderate groups. Raters with six years of study are all in the

very high and high groups. In terms of frequency of MTQ participation, all ratees with participation backgrounds, whether 1-5 times, 6-10 times, more than 10 times, or never participated, are spread predominantly in the moderate group. However, raters with more than 10 times experience also appear to dominate in the very high group. This finding shows that the experience of participating in MTQ is sufficient to show the quality of the ratee's performance in Quranic recitation, especially in the Fashahah aspect.

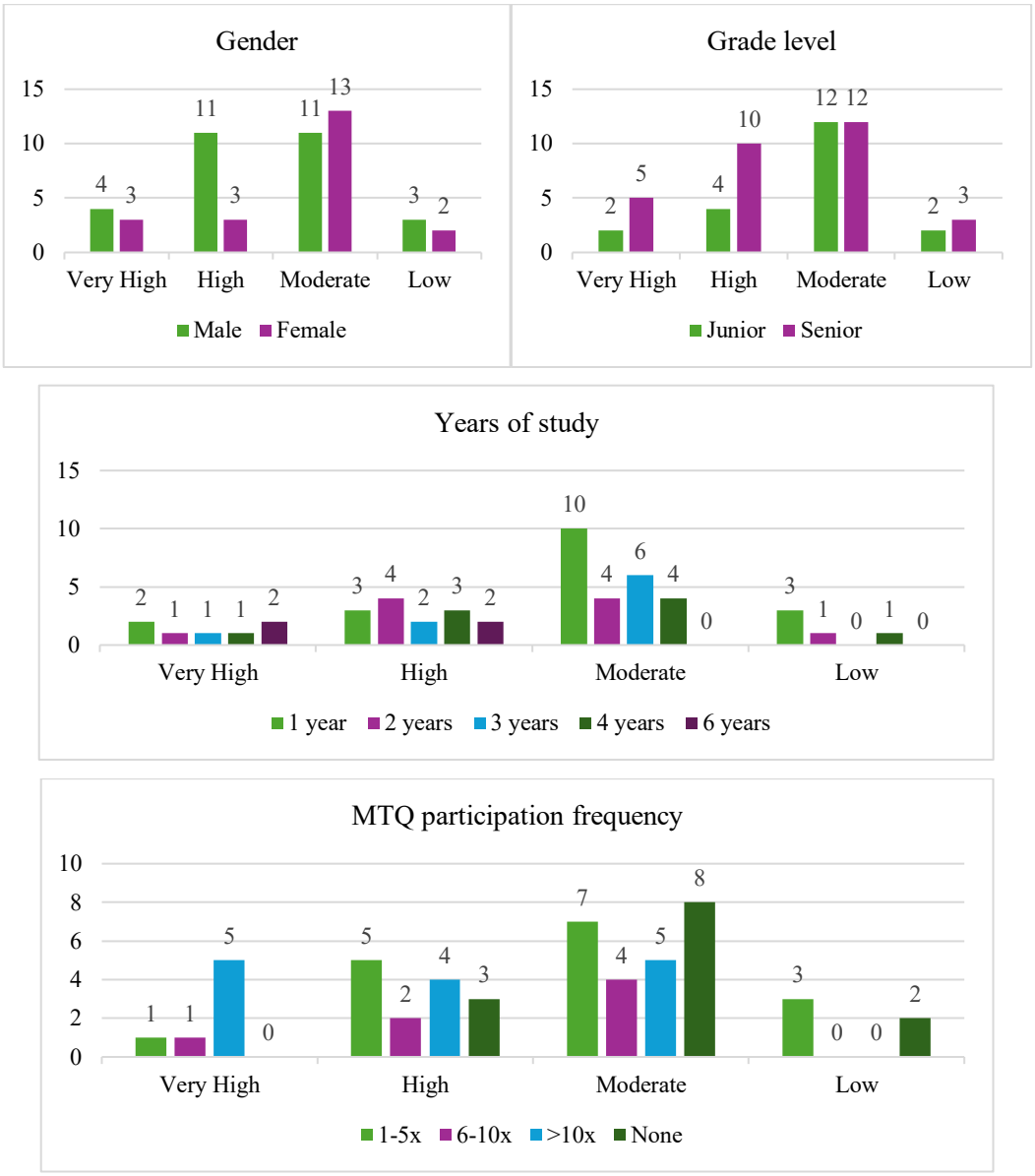
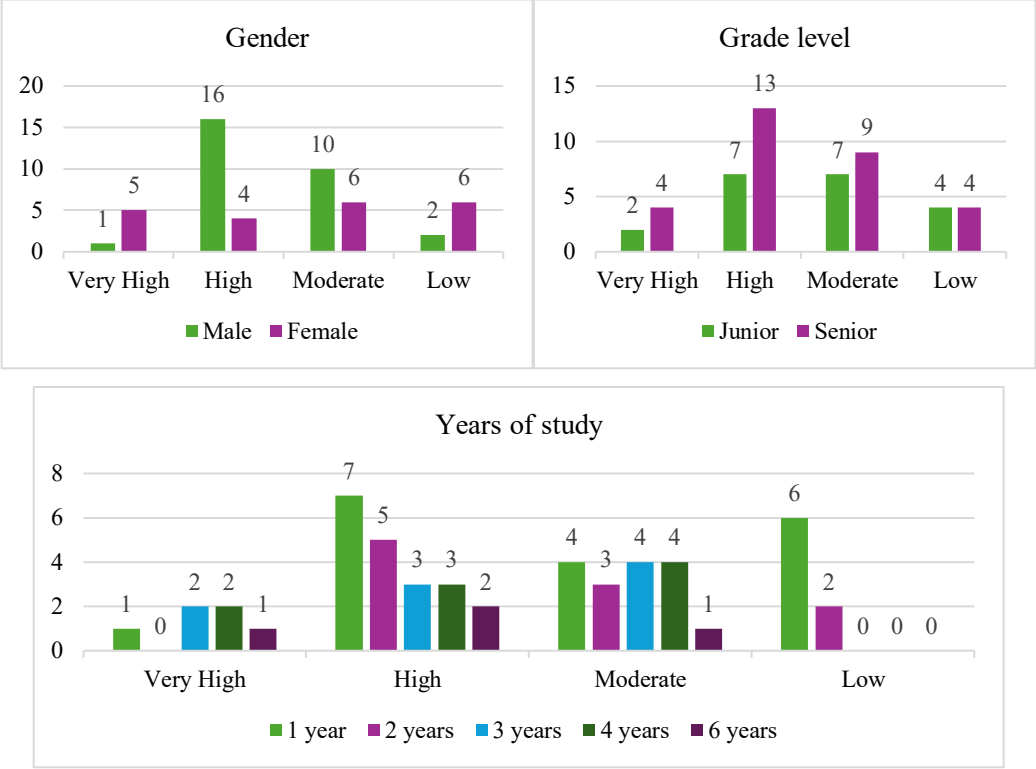


Figure 4. 8 *Ratee distribution across different demographics in the Fashahah dimension*

Lagu

In the *Lagu* dimension, the distribution of ratees considering demographic characteristics can be shown in Figure 4.9. Overall, the distribution of ratees appears to be

quite dominant in the middle groups, namely high and moderate. In terms of gender, male ratees are spread widely in the high group, while female ratees tend to be in the moderate and low groups. In terms of grade level, senior ratees tend to have high abilities in the *Lagu* aspect, while junior ratees tend to have only moderate abilities. As for the aspect of years of study, ratees with a study period of 1, 2, and 6 years are mostly in the high group, while ratees with a study period of 3 and 4 years are in the moderate group. Reports related to the frequency of ratee participation in MTQ show that ratees with 1-5 times experience are spread widely in the moderate group. Ratees with a participation frequency of 6-10 times and more than 10 times tend to be in the high group. Ratees who have never participated in MTQ are spread in the low group. Once again, these results show that experience in MTQ participation also influences the level of ratee ability, especially in the *Lagu* aspect of Quranic recitation.



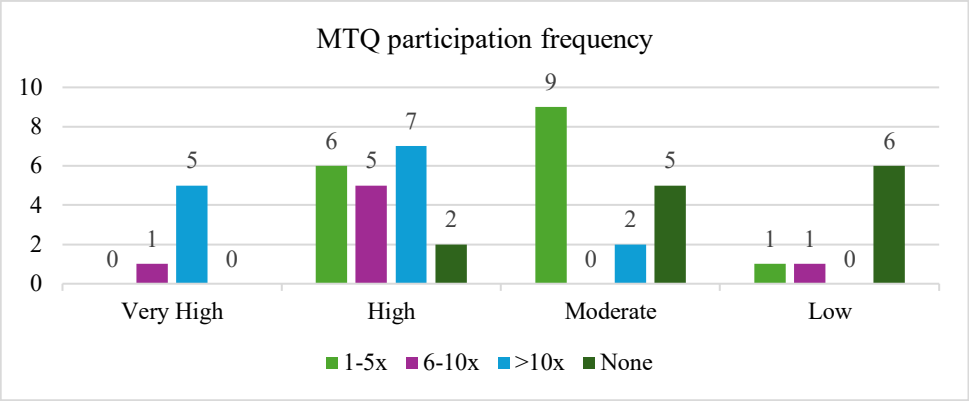


Figure 4. 9 *Ratee distribution across different demographics in the Lagu dimension*

Suara

In contrast to the previous dimensions, in the *Suara* dimension, the distribution of ratee abilities appears quite concentrated in the moderate group (see Figure 4.10). In terms of gender, both males and females are in the moderate group. Interestingly, not a single male ratee is recorded in the low group. Looking at the grade level aspect, both junior and senior ratees are predominantly in the moderate group. In terms of years of study, ratees with 1, 3, and 4 years of study dominate the moderate group. Meanwhile, ratees with 2 years of study are spread across the high and moderate groups. Similarly, ratees with 6 years of study are in the very high and moderate groups. In terms of frequency of MTQ participation, ratees with 1-5 times participation frequency, 6-10 times participation frequency, and those who have never participated are consistently in the moderate group. Meanwhile, ratees with more than 10 times of experience are in the very high group. Interestingly, not a single ratee who has participated in MTQ is in the low group in terms of their vocal ability. This shows that ratees who have participated in MTQ do have good voice quality for Quranic recitation performances.

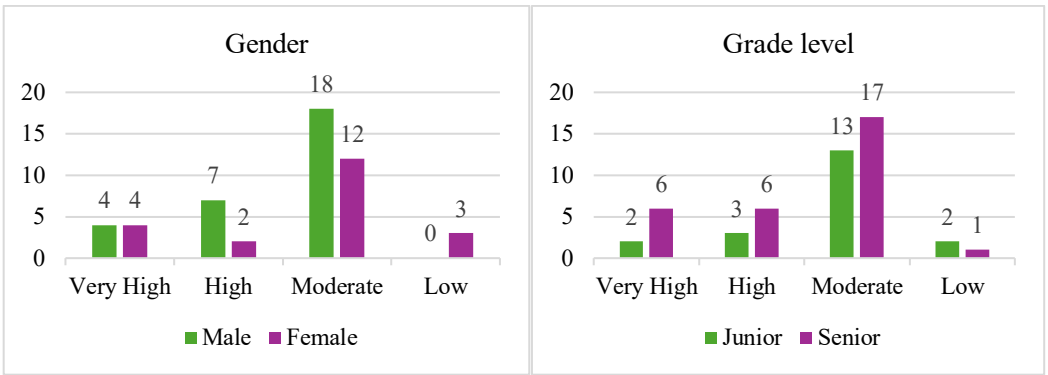


Figure 4. 10 *Ratee distribution across different demographics in the Suara dimension*

Overall, this subsection provides an overview of the demographic distribution of the raters and ratees who participated in this study. In the rater group, significant findings indicate that raters tended to be moderate in their assessments. Meanwhile, in the ratee group, the majority had moderate Quranic recitation abilities. However, the number of ratees in the high and very high-level groups tended to be higher, indicating that ratee abilities were quite high in Quranic recitation. One demographic aspect that appeared most significant to ratee abilities was the frequency of participation in the MTQ competition, particularly for the *Fashahah, Lagu, and Suara* dimensions.

4.2. Findings

The data conditions on all dimensions were presented in the previous section, which shows that the data were in the good category and warrant further analysis. This section aims to assess the quality of the Quranic recitation assessment rubric based on the output of the MFRM analysis. The discussion in this section is structured following the sequence of research questions to ensure a systematic flow of discussion. Each psychometric aspect, validity, reliability, and fairness will be analysed separately by referring to "The Standards

for Educational and Psychological Testing" formulated by AERA, APA, and NCME in 2014.

4.2.1. Validity of the Quranic recitation Assessment Rubric

This section is aimed at addressing the first research question: *To what extent does the Quranic recitation assessment rubric demonstrate psychometric validity based on the results of MFRM analysis?* Referring to The Standards (2014), the analysis will focus on two sources of validity evidence, namely evidence based on internal structure and evidence based on response processes. To assess the validity of the internal structure, an item-level analysis will be conducted to determine the validity through the evidence by evaluating the infit and outfit mean square (MNSQ), infit and outfit Z-standardized (ZSTD), and point-measure correlations (PtMea Corr.) values. It should be noted that the model-data fit analysis in this section will only focus on one facet, which is the items, considering the need is exclusively to see the construct validity of the rubric. For the rater and ratee data fit analysis will be analysed in the next section. Besides that, items ordering through the report of the Wright map, and the unidimensionality test through the variance explained by Rasch measures percentage, will be considered to provide meaningful evidence in determining the validity of the rubric. These indicators will also determine the degree to which each item in each dimension contributes meaningfully to the construct being measured. Furthermore, to assess the validity in terms of evidence based on response processes, the functionalities of the rating scale in each dimension will be analysed. The goal is to determine whether the assessment category scale is used consistently and operates according to the expected assumptions. These results will help in validating the appropriateness of the response process given by the rater. Through these two sources, these analyses can provide empirical evidence to support the psychometric validity of the Quranic recitation assessment rubric.

Evidence of validity based on internal structure

Analysis of the internal structure is performed by evaluating the item fit, item distribution shown by the Wright map, and unidimensionality test report on each dimension of the rubric. Table 4.5 is a combined item report that shows the total score, data count, model measure, standard error, infit and outfit MnSq, and ZSTD, point measure correlation, and predicate fit category for infit and outfit MnSq. Value indicators that require further analysis because they do not meet the ideal standards of data fit are given an underscore label. Figures 4.11 to 4.14 is a set of Wright maps that visualize the distribution of items

within each dimension on a logit scale. This logit scale is the same as the other facets, rater and ratee, and rating scale categories, to show the extent to which the items can cover the range of participants' performance. In addition, the unidimensionality report was also performed to demonstrate whether the items can measure the full range of the respondent's ability while assessing a single variable comprehensively.

Table 4. 6 *Item fits statistics across dimensions*

Item	Total score	Total count	Model		MnSq		MnSq fit category		ZSTD		Cor.
			Mea	S.E.	Infit	Outfit	In	Out	Infit	Outfit	PtMea
T1	792	200	-0.30	0.09	0.78	0.74	A	A	-2.1	-2.5	0.56
T2	714	200	0.20	0.08	0.68	0.67	A	A	-3.6	-3.5	0.49
T3	783	200	-0.24	0.08	1.05	1.04	A	A	0.4	0.4	0.55
T4	689	200	0.34	0.07	1.44	1.46	A	A	4.0	4.0	0.47
Mean	744.5	200	0.00	0.08	**0.98**	**0.98**	**A**	**A**	**-0.3**	**-0.4**	**0.52**
F1	871	192	0.85	0.09	0.96	0.76	A	A	-0.1	-1.0	0.42
F2	789	192	1.07	0.08	1.11	1.07	A	A	0.9	0.5	0.49
F3	546	192	0.29	0.16	0.98	0.75	A	A	0.0	-0.6	0.28
F4	573	192	-2.22	0.58	0.94	0.32	A	B	0.0	-0.9	0.21
F5	No Detected										
Mean	694.8	192	0.00	0.23	**1.00**	**0.73**	**A**	**A**	**0.2**	**-0.5**	**0.35**
L1	982	200	-2.31	0.26	1.54	0.67	C	A	2.0	-0.5	0.38
L2	957	200	-1.22	0.18	1.75	0.80	C	A	3.7	-0.6	0.48
L3	713	200	1.96	0.09	1.13	1.12	A	A	1.2	1.1	0.64

L4	763	200	1.54	0.09	0.94	0.96	A	A	-0.5	-0.3	0.65
L5	897	200	0.03	0.12	0.97	0.84	A	A	-0.1	-0.9	0.64
Mean	862	200	0.00	0.15	**1.27**	**0.88**	**A**	**A**	**1.3**	**-0.3**	**0.56**
S1	437	200	0.70	0.11	0.80	0.77	A	A	-2.5	-2.0	0.63
S2	513	200	-0.33	0.13	1.01	1.01	A	A	0.1	0.1	0.41
S3	534	200	-0.70	0.14	1.12	0.93	A	A	-0.2	-0.2	0.36
S4	552	200	-1.09	0.16	1.21	1.13	A	A	0.5	0.5	0.28
S5	379	200	1.42	0.11	1.00	1.03	A	A	0.3	0.3	0.56
Mean	483	200	0.00	0.13	**1.03**	**0.97**	**A**	**A**	**0.0**	**-0.3**	**0.45**

Note: T: *Tajwed* dimension; F: *Fashahah* dimension; L: *Lagu* dimension; S: *Suara* dimension. The meaning of the fit category (A, B, C) is based on Table 3.8.

The analysis of the internal structure validity of the Quranic recitation assessment rubric is conducted by evaluating the **item-fit statistics** based on the MFRM model. Three main indicators recommended by Boone et al. (2014) will be the guideline for assessing item fit, namely: the infit and outfit mean square (MnSq) values must be in the range of 0.5 to 1.5, the infit and outfit Z standardized (ZSTD) values should ideally be in the range of -2.0 to +2.0, and the point measure correlation (PtMea Corr.) must be positive because it shows the relationship between item responses and the traits being measured (Boone et al., 2014). In addition, the fit category predicate will also be considered by referring to the Table 3.8 framework as suggested by Engelhard and Wind, (2018). Meanwhile, the interpretation of ZSTD will refer to the Linacre (2002) guidelines as shown in Table 3.9.

Overall, the item in all dimensions is considered to have a good fit because they are within the ideal range (see the mean of fit statistics for each dimension). These results suggest that the rubric structure is well-designed, and that the items used are consistent with the assumptions of MFRM model. However, if the analysis further to the item level, several items need to be criticized because they show deviations from the ideal criteria. *First*, the *Tajwed* dimension has three items, namely T1, T2, and T4. These three items show a fit category that is still acceptable, with 100% of items getting the A-productive for measurement category, when referring to the infit and outfit MnSq values (Engelhard & Wind, 2018). However, a significant deviation occurred because the infit and outfit ZSTD

values exceeded the extreme limits (T1: -2.1 and -2.5; T2: -3.6 and -3.5; T4: 4.0 and 4.0). This indicating the rater responses to these items is misfit with the model. Several factors can be the cause, such as in items T1 and T2, the infit and outfit ZSTD values are below the standard which indicating evidence of overfit, meaning the response pattern given by the rater is indicated as too consistent or too predictable; while in item T4 which has a ZSTD value above the standard indicating evidence of misfit, meaning the response pattern tends to be too random or very unexpected (Linacre, 2002). This situation encourages researchers to question again whether the instructions or definitions of the scope of interpretation of these items are aligned with the intended construct.

Second, items F4 and F5 in the *Fashahah* dimension. Item F4 is recorded as having an outfit MnSq value that is below standard. Referring to the explanation of Engelhard and Wind (2018), this item is considered less productive for measurement but not distorting of measures. This is because the response pattern given to this item is considered too easy to predict or does not provide a new information on the ratee's ability due to the evidence of overfit. The results indicates that item F4 has low discrimination capacity and can be categorized as a "redundant" item. On the other hand, an interesting finding occurred in item F5 where the data provided could not be detected by the model. Referring to Table 4.6, there were indeed invalid responses that were not included in the analysis. The researcher conducted an internal search of the dummy data created and found that the responses given by the rater to item F5 to all ratees were the same, by giving the maximum score. This finding explains that the response to item F5 was not processed because they were uniform and did not provide valuable information for the analysis.

If item F5 is analysed further, it will be found that it is *tamam al-waqt,* which assesses the punctuality of the ratees in Quranic recitation. Referring to the "Guidelines for Musabaqah Al-Quran and Hadith in 2023" (2023), this item exists because there are rules regarding the performance time limit, which is seven to ten minutes. Every ratee who exceeds the time limit will get a point reduction which will affect the *Fashahah* dimension score. Even so, it is odd that items related to performance time are placed in the *Fashahah* dimension, which assesses the accuracy of Quranic recitation. The *Fashahah* dimension related to the theory and practice of recitation. Performance time is another aspect that might be more suitable in the *Lagu* dimension. There, more relevant aspect exists, such as item L1: assessing the first and closing songs, and item L2: number of songs/compositions. This study offers an opinion; this dimension is more suitable for pairing with the aspect of assessing performance time. Regardless, this condition shows that all ratees have displayed

timely Quranic recitation performance as standardized, although indirectly assuming that item F5 is not effective enough in measuring ratee performance in the context of this study. This study assume that this is quite reasonable, because the ratees have received the same courses in the *Pesantren* so that familiarization with the competition is very likely to have occurred.

Third, items L1 and L2 in the *Lagu* dimension simultaneously show evidence of misfit. Item L1 recorded MnSq and ZSTD infit values exceeding the standard (1.54 and 2.0, respectively). Item L2 also recorded MnSq infit and ZSTD infit above the standard (1.75 and 3.7, respectively). The infit of MnSq results that are above the standard indicate that the item provides unproductive measurements, although not distorting measures (Engelhard & Wind, 2018); while the infit of ZSTD that is not achieved explains that the data is unpredictable (Linacre, 2002). This condition indicates too much random variation, suggesting the possibility of multiple interpretations of the scale rubric criteria or that the rater's perceptions are not yet consistent. *Fourth*, item S1 in the *Suara* dimension exhibits overfit evidence in both the infit and outfit ZSTD. This result indicates that the response to item S1 is considered too consistent or too predictable, so it is likely not able to provide new information. This situation can indicate that there is too much uniformity in rater assessments or even an indication of assessment bias.

The next analysis is evaluating the internal structure of the rubric by analysing the order of the items through the **Wright map.** Figures 4.11 to 4.14 are a comprehensive visualization of the ratee, item, and rater distribution in a uniform logit scale. According to Engelhard and Wind (2018), the logit scale used in Rasch measurement allows direct comparison between the three facets. However, the analysis in this part will be focused on the distribution of the items in each dimension. To provide granular information, the value of the mean and standard deviation will be used to categorize the level of item difficulty in each dimension. By referring to the output table from Facet software, Table 7.3.1, the order of the items in each dimension will be determined as good if their distribution is spread across each group difficulty level (Boone et al., 2014).

In the *Tajwed* dimension, looking at the Wright map display in Figure 4.11, the distribution of these four items seemed quite overlapping or appeared to have similar difficulty levels. However, through the Rasch analysis, this detailed information can be identified, providing useful insights for researchers, especially in understanding that the distribution of item difficulty levels in this dimension was good. According to the MFRM analysis reported, the mean of the four items was at a logit of 0.00 with a standard deviation

of 0.28. Through these results, the distribution of item orders was recorded as being good because it had filled all difficulty level groups. The item with a logit value above 0.28 or the one considered the most difficult was item T4. The item with a difficult level that was in the logit range of 0.00 and 0.28 was item T2. Meanwhile, the item with a moderate level of difficulty was item T3, which was in the logit range of -0.28 to 0.00. The item with the lowest difficulty level was item T1, which had a logit value below -0.28. This finding indicates that the item difficulty level in the *Tajwed* dimension is appropriate, as it shows a distribution of difficulty that matches the target population, although the discrimination among items can still be improved (Engelhard & Wind, 2018).

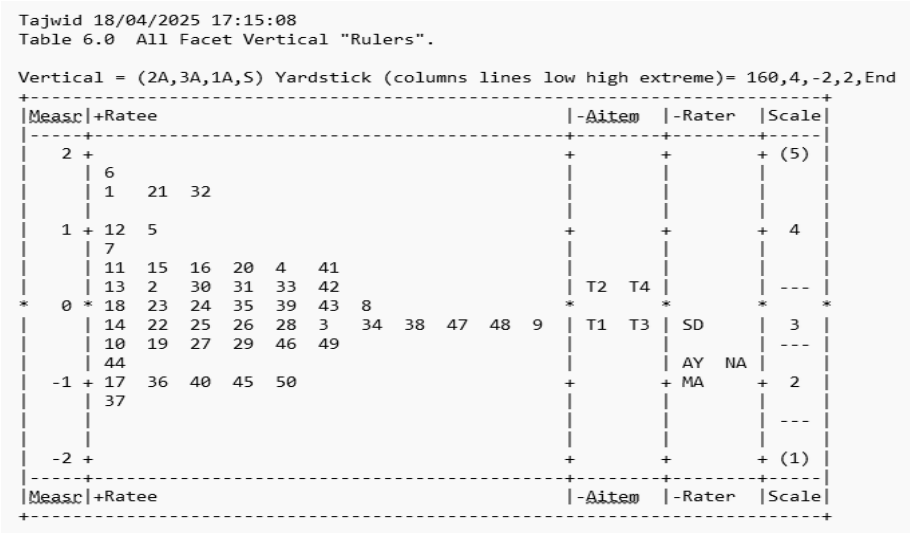


Figure 4. 11 *Wright map of Tajwed dimension*

Figure 4.12 displays the distribution of items in the *Fashahah* dimension. In this figure, there are slight differences, especially in the item scale column, because the model analysis uses a partial credit model (Engelhard, 2013; Masters, 2016; Andrich, 2019). Overall, the item difficulty appears quite good, being evenly distributed from the highest to the lowest logit values, although one item is significantly different from the others. However, considering the mean value of 0.00 and standard deviation of 1.31, the distribution of item difficulty in this dimension is uneven. Of the five items used, items F1, F2, and F3 are considered to have the same difficulty level, which is at a high level. Meanwhile, item F4 is considered to have a low difficulty by the model. Furthermore, item F5, which exhibits a uniform scoring pattern, was not detected in the model's analysis estimation. These results demonstrate the need for item evaluation to ensure differential discrimination across all difficulty levels.

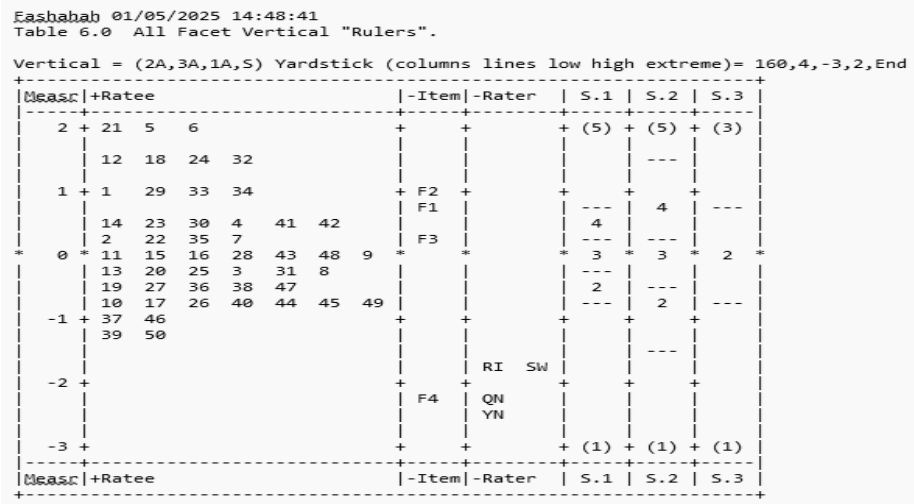


Figure 4. 12 *Wright map of Fashahah dimension*

In Figure 4.13, the item distribution shown on the Wright map appears very good. This is in line with the information on the distribution of item difficulty levels, which refers to the mean logit value of the five items, which is at 0.00 with a standard deviation of 1.61. Item L3 is recorded as having a logit value above 1.61, explaining that this item is included as an item with the highest level of difficulty. Followed by items L4 and L5, which are in the range of 0.00 to 1.61 and considered as the items with a high level of difficulty. The item with a moderate level of difficulty is item L2 with a logit value between 0.00 and -1.61. Meanwhile, the item with the lowest level of difficulty is item L1, which has a logit value below -1.61. Through these results, it can be understood that the distribution of item difficulty discrimination in the *Lagu* dimension is considered very good.

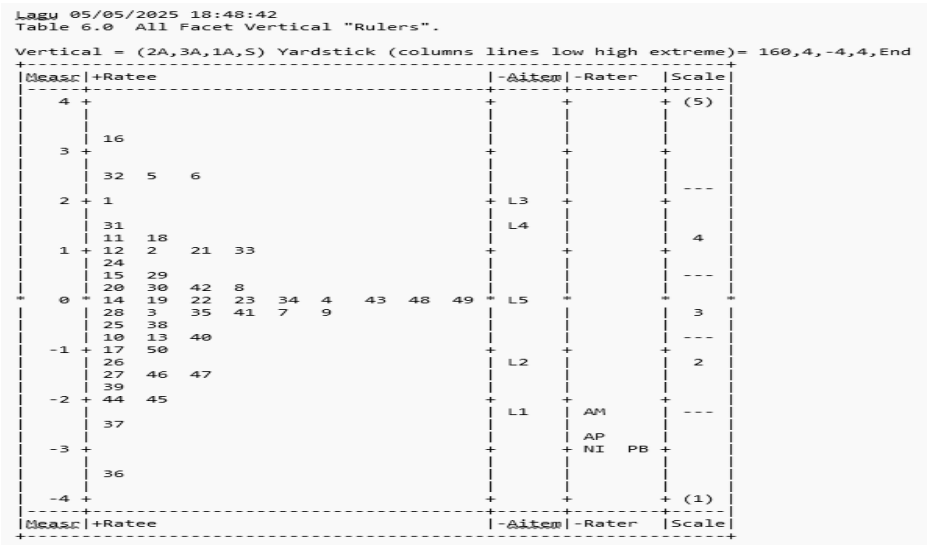


Figure 4. 13 *Wright map of Lagu dimension*

Figure 4.14 shows the distribution of the five items in the *Suara* dimension with good discrimination. None of the items appear to be overlapping. It can be assumed that all difficulty levels are covered by the items in this dimension. Referring to the overall mean value of the items, which is at a logit of 0.00 with a standard deviation of 0.93, it explains that the item with the highest level of difficulty is item S5. This is followed by item S1, which is considered one of the difficult items because it has a logit value between 0.00 and 0.93. Meanwhile, items with a moderate level of difficulty are items S2 and S3, both of which are in the logit range of 0.00 to -0.93. The item with the lowest level of difficulty is item S4, because it has a logit value below the standard deviation (-0.93). These results indicate that, in addition to the distribution of items in the *Lagu* dimension being assessed as very good, the items in the *Suara* dimension are also assessed as having even difficulty discrimination by the model. In sum, all item orders, based on the analysis of the level of discrimination and difficulty, showed that the items across all dimensions were well distributed. Except for the items in the *Fashahah* dimension, which needed treatment to further expand discrimination across all levels.

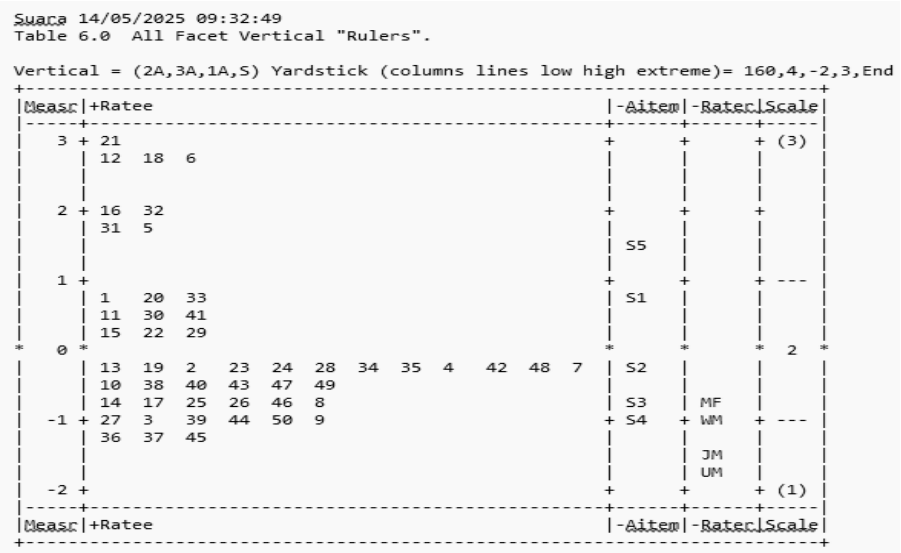


Figure 4. 14 *Wright map of Suara dimension*

In addition to providing analysis results on internal structure, other important indicators must be revealed. **Unidimensionality** is an important component that ensures the items measure only one variable (The Standards, 2014). The raw score variance explained by the Rasch measure is an indicator that can provide evidence of validity (Wind, 2023). The results of this analysis have been presented in section 4.1.2, but only separately.

For this reason, Figure 4.15 is provided to show a comparison between dimensions to provide a comprehensive understanding.

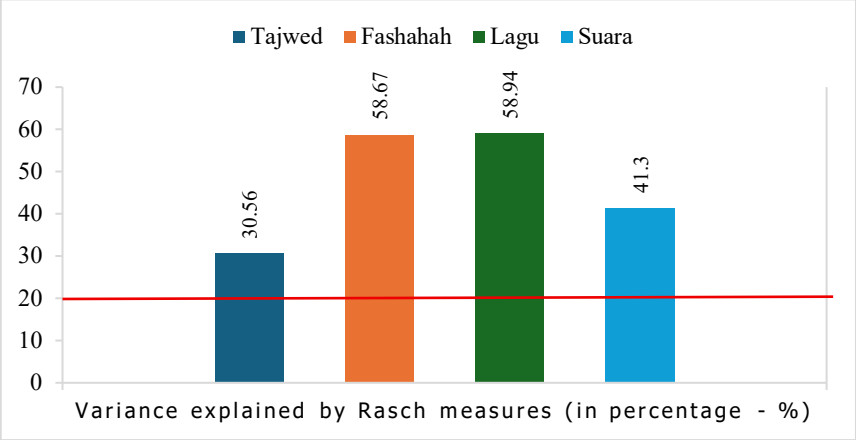


Figure 4. 15 *The variance explained by measures across dimensions*

In general, the percentage of variance explained by the Rasch measure has met the minimum standard (> 20%) recommended by Reckase (1979). This confirms that all dimensions have items that measure only one construct or unidimensionality. This provides good evidence that, from the perspective of internal structure, all variance in the data can be explained in one latent dimension, following the model assumptions, so that all four dimensions, although the percentage is not too high.

Overall, most items provide a good fit category with the model, although several problematic items recommend the need for improvement in terms of the formulation and operationalization of rubric criteria. The order of the items in terms of their discrimination on the level of difficulty shows good separation, even though a treatment in the *Fashahah* dimension is recommended to be conducted. The examination of unidimensionality also reveals that the items only measure one construct (dimension). However, the deviation of fit values found in several items can be the basis for reviewing the assessment indicator descriptors, construct coverage, and rater training need to be done to create a more valid rubric. These findings provide evidence-based support in revealing the validity of the internal structure of the rubric, as well as providing a direction for improving the accuracy and representativeness of Quranic recitation assessment.

Evidence of validity based on response processes

Other evidence of validity is provided by evaluating the response processes or function of the assessment rating scale in each dimension. For the *Tajwed, Lagu,* and *Suara* dimensions, category statistics reports are presented in aggregate per dimension in Table 4.7 because it has a uniform rating scale between items (Wind, 2023). Meanwhile, for the *Fashahah* dimension, it is reported separately in Table 4.8 because the use of scales varies between items. The separation is done because the rating scale analysis method is a bit different. According to Wind (2023), which refers to researchers Engelhard and Wind (2013, 2018), Linacre (2002), Wright and Masters (1982), the dimensions in Table 4.6 needs to be analysed with the Rating Scale Model (RSM), while specifically for *Fashahah* dimension in Table 4.7 needs to be analysed with the Partial Credit Model (PCM) (Wind, 2023). For that, referring to Wind (2023), categories ordering, categories precision, and category comparability (specifically PCM), will be the guidelines of analysis to provide evidence of validity on response processes.

Table 4. 7 *Category statistics report for items in Tajwed, Lagu, and Suara dimension*

Scale cat.	**Counts used**	**(%)**	**Avrg. mea.**	**Outfit MnSq**	**Rasch Andrich threshold**		**Adjacent threshold distance**	**Cat. peak prob.**
					Mea.	**S.E.**		
Tajwed dimension								
1	42	5%	-0.13	1.0	N/A	N/A	N/A	100%
2	57	7%	-0.01	0.9	-0.44	0.17	N/A	24%
3	174	22%	0.33	0.9	-0.95	0.12	-0.51	34%
4	335	42%	0.75	1.0	-0.13	0.08	0.82	51%
5	192	24%	1.18	1.0	1.52	0.09	1.65	100%
Lagu dimension								
1	21	2%	-0.83	1.6	N/A	N/A	N/A	100%
2	44	4%	-0.4	1.8	-1.51	0.27	N/A	40%
3	109	11%	0.60	1.2	-0.76	0.17	0.75	42%
4	254	25%	1.50	0.5	0.30	0.12	1.06	52%
5	572	57%	4.04	0.9	1.97	0.09	1.67	100%
Suara dimension								
1	148	15	-0.27	1.0	N/A	N/A	N/A	100%

2	289	29	0.47	0.9	-0.59	0.11	N/A	47%
3	563	55	1.99	1.0	0.59	0.08	1.18	100%

The first analysis of Table 4.7 focuses on the category ordering aspect. Referring to Wind (2023), there are three ways that can be used to evaluate this aspect, examining the average measure value, the Rasch Andrich threshold measure, and the order of the rating scale in the probability curves. In general, all dimensions have a logically structured category order. A low scale is used to provide an assessment for ratees with poor performance, while a high scale is used to provide an assessment for excellent performers. However, in the *Tajwed* aspect, an irregularity in the threshold was found between scale categories 2 and 3 (-0.4 and -0.95, respectively). According to Tennant (2004), this indicates a disordered threshold, meaning the scale was inappropriately used in category 3 by the rater. Meanwhile, the category order is consistently shown in the *Lagu* and *Suara* dimensions, which explains that the categories used meet the criteria for ordered thresholds because they increase consistently and systematically. Visually, the graphical display of the order of categories across dimensions can be seen in the probability curves, Figure 4.16.

The next analysis that can be done by referring to Table 4.7 is the aspect of rating scale category precision. This evaluation is important to help the researcher explain how well each category works in identifying latent factors in the performance of ratees at various levels of ability. Wind (2023) states that there are three ways to do this: identify the distance between thresholds, assess the model data fit in each category, and distinguish categories on probability curves. Overall, the precision results in these three dimensions require improvement in several parts, especially in the middle scale. This argument was drawn after several adjacent threshold distances in all dimensions failed to meet the minimum standards recommended by Linacre (2002). According to Linacre, the distance between thresholds must be less than five logits and more than the minimum standard; the minimum standard for the threshold distance is determined by the number of categories used. Through the following formula, $\mathrm{In}\left(\frac{x}{(m-x)+1}\right)$, where *In* is for natural logarithm, *x* is for category number, *m* is for the number of dichotomies (Linacre, 2002), the researcher determined that the minimum distance for the rating scale categories 1-5 is 0.98 logit, while for categories 1-3 it is 1.39 logit. Based on the results of the adjacent threshold distance in the *Tajwed* dimension, two out of three distances did not meet the minimum standard. These results indicate that raters tend to have difficulty in distinguishing categories 2 to 3 and 3 to 4. According to Wind (2023), this can reduce the level of measurement precision because it

provides less information about the ratee's performance. In the *Lagu* and *Suara* dimensions, the threshold distance for categories 2 to 3 (*Lagu*: 0.75 < 0.98; *Suara*: 1.18 < 1.39, respectively) also does not meet the minimum standard and therefore does not meet the category precision criteria.

Meanwhile, about the fit categories model, the Outfit MnSq indicator is used according to general standards: the closer the value is to 1.0, the more ideal (Boone et al., 2014; Smith, 2004). Table 4.7 shows that almost all categories in the three dimensions are considered a good fit, and some categories are found to have an ideal value. However, categories 1 and 2 in the *Lagu* dimension have outfit MnSq values outside the range of standards, 1.6 and 1.8, respectively. This indicates that these category scales are a misfit.

The results of the precision analysis of each category are clarified by the reports in other columns. Starting with the disordered threshold in the *Tajwed* dimension, the percentage of category use in the assessment is not distributed evenly. The indications of misfit in categories 1 and 2 in the *Lagu* dimension are clarified by the percentage of scale use, which is also too lopsided. Likewise, the distribution of category 3 in the *Suara* dimension has a fairly large percentage when compared to the others, although it still categorizes as quite balanced. Visualization of this condition, which clarifies the differences in distribution of each category, can also be seen in Figure 4.16, which depicts the probability curves of the three dimensions. It should be noted that even though there are several category scale transitions below the minimum standard, the distribution visualization in Figure 4.16 still shows differences between one category and another, though it must be acknowledged that the distribution is too narrow.

Table 4. 8 *Category statistics report for each item in the Fashahah dimension*

Scale cat.	**Counts used**	**(%)**	**Avrg. mea.**	**Outfit MnSq**	**Rasch Andrich threshold**		**Adjacent threshold distance**	**Cat. peak prob.**
					Mea.	**S.E.**		
Item F1								
1	7	4%	0.09	0.4	N/A	N/A	N/A	100%
2	4	2%	0.47	0.6	0.90	0.42	N/A	14%
3	8	4%	0.60	0.6	-0.15	0.35	1.05	13%
4	33	18%	1.05	1.0	-0.60	0.27	0,45	27%

5	128	71%	1.46	1.0	-0.15	0.18	0.45	100%
Item F2								
1	6	3%	0.05	0.8	N/A	N/A	N/A	100%
2	11	6%	0.33	0.9	-0.38	0.44	N/A	27%
3	35	19%	0.67	1.0	-0.70	0.28	0.32	36%
4	44	24%	1.08	1.4	0.54	0.19	1.24	30%
5	84	47%	1.35	1.2	0.54	0.17	0.00	100%
Item F3								
1	9	5%	0.89	0.6	N/A	N/A	N/A	100%
2	12	7%	1.56	0.9	0.95	0.36	N/A	16%
3	159	88%	1.90	1.0	-0.95	0.24	1.90	100%
Item F4								
1	No result							
2	3	2%	2.89	0.3	N/A	N/A	N/A	N/A
3	177	98%	4.36	1.0	N/A	N/A	N/A	N/A
Item F5 is not reported								

A similar analysis is also given to the *Fashahah* dimension, and its statistical indicator results are reported separately in Table 4.8. Unlike the response process analysis of other dimensions, one additional category besides category ordering and category precision will be analysed: category comparability (Wind, 2023). Category comparability analysis is a comparative analysis of the validity of the rating scale order and precision of items under unique conditions (Wind, 2023). Before that, it should be noted that item F5 does not have a statistical report due to the data is invalid.

In terms of category order, all items exhibit orderliness from the lowest to the highest category when referring to the average measure value. However, this differs from the report given by Rasch Andrich threshold measures in that there are some thresholds not in order in all items. In item F1, categories 2 and 3 move down from what should have increased regularly. In item F2, deviations also occur in category 3 (-0.70). Item F3 also shows a threshold value that deviates from category 3. Unlike the others, item F4 does not have a threshold report because no data was provided for category 1; so that, threshold analysis cannot be performed. The probability curves also visually show the scale distribution for each item and further clarify its regularity and disorder threshold conditions (Figure 4.17). These results show that, in terms of in-order regularity, all items provide monotonic results.

However, the disordered threshold condition indicates a dysfunction in the category transition that needs to be evaluated.

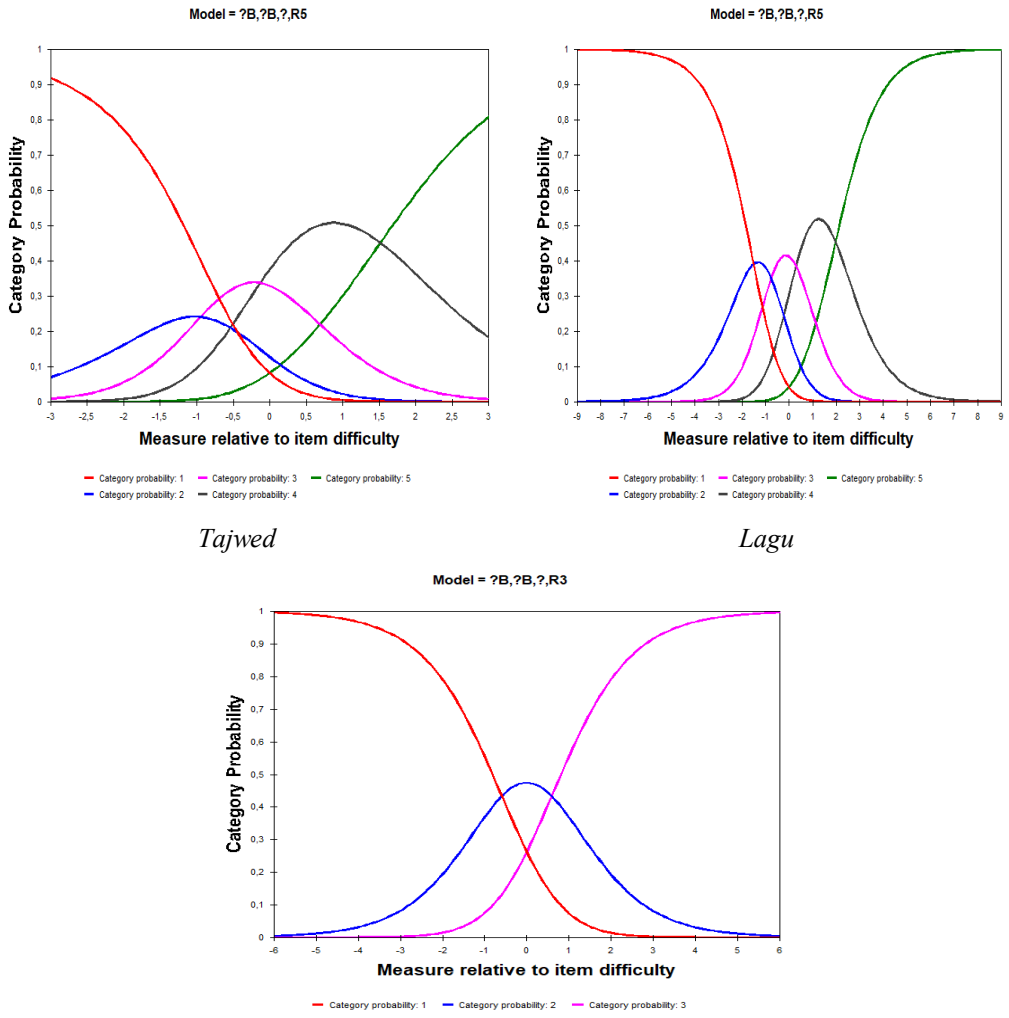
In terms of category precision, several indicators can be considered (Wind, 2023). According to the distance between threshold results, all categories must meet the minimum standard according to Linacre's (2002) recommendation. Rating scale 1-3 must be more than 1.39 logits, and rating scale 1-5 must be more than 0.98 logits. Overall, only item F3 has the most acceptable distance threshold. Meanwhile, items F1 and F2 have two category transitions that do not meet the minimum standard (F1: distance from category 3 to 4 and 4 to 5; F2: distance from category 2 to 3 and 4 to 5, respectively). These results indicate that the transition distance between scales is too narrow to provide new information about ratee performance. Meanwhile, item F4 does not have an adjacent threshold distance because there is no report on the Rasch Andrich threshold. Considering the Outfit MnSq value, only two scale categories have values below the range and are indicated as overfit. These categories are scale 1 on item F1 and scale 2 on item F4, showing that the responses on both scales are too predictable or have low variation.

In terms of category comparability, the conditions of items F1 and F2 are slightly different from those of items F3 and F4. Items F1 and F2 have problems with category functionality and distance thresholds that are too small for several categories transition. In contrast, item F3 has a distance threshold that meets the standard, but with a few problems in the order category. As for item F4, the effectiveness of the scale that only occurs in two categories explains that the scales in this item are less than optimal because it does not utilize the entire range of categories. To provide clearer information, Figure 4.17 provides a visual depiction of the differences in the distribution of each category in each item.

Analysis of the validity evidence reveals that validity based on the response process indicates the need to improve the precision of the assessment of the scale category functionality. While most have met the category ordering criteria consistently and systematically, the threshold distance between categories in several dimensions is still too narrow. This certainly impacts the level of assessment precision, ultimately affecting the validity of the assessment. Therefore, improvements to the rubric regarding the scale category aspect are needed, particularly when compiling descriptors or selecting the rating scale range. The reason is that many findings show that the scales provided have not been fully distributed in a balanced manner.

Overall, the results of the analysis on the internal structure and response patterns

have provided answers to the first question in this study. In general, the Quranic recitation assessment rubric is considered valid, with some notes. This is evidenced by the strong internal structure and the pattern of the rater response process that understands the distribution of the order of each scale category. However, in terms of item suitability, several items still need improvement. Likewise, in terms of the precision aspect of the scale category functionality, revisions need to be made so that the assessment given can increase the validity of the Quranic recitation assessment.



Tajwed *Lagu*

Suara

Figure 4. 16 *The order of probability curves in the Tajwed, Lagu, and Suara dimensions*

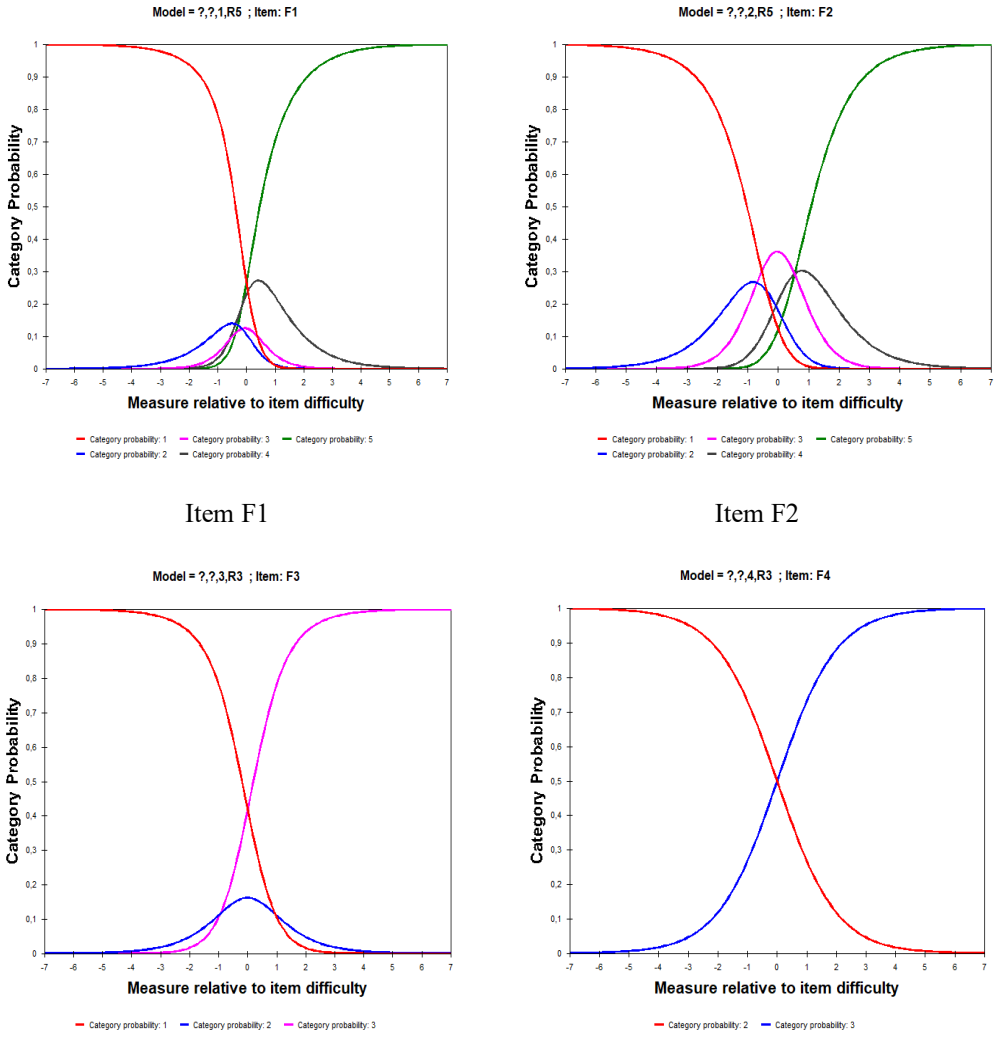


Figure 4. 17 *The order of probability curves for each item in the Fashahah dimensions*

4.2.2. Reliability of the Quranic Recitation Assessment Rubric

Within the framework of "The Standards for Educational and Psychological Testing" formulated by AERA, APA, and NCME, reliability is defined as the extent to which a measurement component is reliable and consistent across different types of conditions. There are three terms used to describe reliability: stability, equivalency, and internal consistency (Sumintono & Widhiarso, 2015). This is the basis that this study used to formulate the second research question: *How reliable is the rubric in ensuring consistent scoring across different judges and participants, as evidenced by MFRM analysis findings?* To answer this question, Engelhard and Wang (2021) provide several sources of evidence that can be found by referring to the eight interrelated clusters outlined by The Standards (2014). However, since this section is focused on revealing the consistency and precision

of rubric reliability, only three clusters are recommended to use as evidence: reliability and generalizability coefficients, standard errors of measurement, and decision consistency (The Standards, 2014). The Standards (2014) state that evidence for these three clusters can be seen in several indicators, including reliability of person scores, precision of person measures, reliability of item scores, precision of item calibrations, and all the evidence related to the consistency and precision indicators, such as the inter-rater agreement and root mean square error. In this section, this study will utilize all the related reports from the MFRM output to provide strong evidence in assessing the reliability of the Quranic recitation assessment rubric. The details of the analysis will be followed according to the dimension groups.

Tajwed dimension

Table 4. 9 *Ratee, Item, and Rater measurement report in Tajwed dimension*

	Ratee	**Item**	**Rater**
N	50	4	4
Root Mean Square Error (RMSE)	0.30	0.08	0.08
Adjusted (True) Standard Deviation (SD)	0.61	0.26	0.21
Separation Ratio	2.05	3.30	2.61
Strata	3.06	4.73	3.82
Reliability Index	0.81	0.92	0.87
Model fixed chi-square	233.4	47.9	33.3
Significance (Probability) (*p*)	0.00	0.00	0.00
Degree of freedom (d.f.)	49	3	3
Model random effect chi-square	39.8	2.8	2.8
Significance (Probability) (*p*)	0.00	0.00	0.00
Degree of freedom (d.f.)	48	2	2
Inter-Rater agreement opportunities	-	-	1200
Observed Exact Agreements (%)	-	-	37.6%
Expected Agreements (%)	-	-	34.0%

In providing the evidence for the consistency and precision of the *Tajwed* dimension, Table 4.9 explains the details of their statistical summary. ***First,*** **Ratee**. Based on the results of the MFRM analysis, the data from the assessment showed good conditions in

differentiating the ratees' ability in reciting the Quran. The value of Root Mean Square Error (RMSE) is 0.30, which indicates that the level of measurement error is low because the estimation of the ratee's ability level is quite accurate. Linacre (2023) stated that a low RMSE value indicates the stability of the logit estimate in the MFRM model, thus allowing a more reliable evaluation of the Quranic recitation performance. Apart from that, the adjusted (true) standard deviation (SD) value of 0.61 shows that there is a wide enough distribution. Therefore, the assessment process in the *Tajwed* dimension can identify ratee's abilities variations in reciting the Quran (Engelhard & Wang, 2024).

The ratee separation ratio is at 2.05, indicating that the rubric possesses the capability to discern three disparate performance levels (Bond et al., 2021). According to the formula H = [(4 x G) + 1] / 3, the value produces the strata of 3.06, which is rounded up to produce the decision to divide the ratee into three groups of level performers (Bond et al., 2021; Bond & Fox, 2015). The division of these three groups confirms the existence of a strong discriminatory function of the rubric in classifying ratees based on their abilities. The reliability value of 0.81 confirms that the rubric on the *Tajwed* dimension is consistent in assessing the Quranic recitation performance of ratees (Engelhard & Wang, 2024; Fisher, 2007). Additionally, because the reliability value has met the minimum reliability standard (Fisher, 2007), the rubric on the *Tajwed* dimension can be accepted, especially for evaluative and diagnostic purposes.

The chi-square value of the fixed model is 233.4 with a degree of freedom of 49 and a significance probability (p) = 0.00. According to Linacre (2023), this value indicates that the assumption that all participants have the same ability does not follow the data. It means there is a significant difference in the ratee that can be detected through the rubric, specifically in the *Tajwed* dimension. In addition, the random effects model, which produces a chi-square value with degree of freedom 48 and p value = 0.79, shows that the data statistically corresponds to a normal distribution as assumed by the model.

***Second,* Item.** To explain the condition of the items in the *Tajwed* dimension, Facets software produces several statistical reports on items as presented in Table 4.9. In general, the quality of the items used in the *Tajwed* dimension is in the very good category. This is based on a low RMSE value of 0.08, indicating that the estimated item difficulty level used, referring to Linacre's (2023) explanation, has a very small error rate. The adjusted S.D. of 0.26 also explains that the level of difficulty varies between items. This means that each item has a good variety of levels of difficulty in grouping participants' abilities fairly

because the items used do not have a homogeneous difficulty standard (Engelhard & Wang, 2024).

The findings above were then confirmed by the separation ratio value, which was 3.30. This number shows that the items can be distinguished at least into three significantly different groups. The high strata value, which is 4.73, also explains that the items in the Quranic recitation assessment rubric on the *Tajwed* dimension can even be classified into four consistent difficulty groups. Referring to the explanation of Bond et al. (2021), the higher the separation and strata values, the better the ability of the assessment instrument to distribute the level of difficulty between items. In addition, the reliability index, which is at 0.92, also shows that all items on the *Tajwed* dimension are very reliable even in replication measurements because of their high level of reliability (Fisher, 2007).

The results of the fixed effect model test show that the chi-square value is at 47.9 with a p value = 0.00 with a degree of freedom (d.f.) at 3. These numbers explain that the assumption of all items having homogeneity in difficulty levels is statistically rejected (Linacre, 2023). Testing on the random effect model also confirms that there is no significant difference between empirical data and the assumption of normal distribution in the Rasch model (Linacre, 2023). This is indicated by the chi-square value, which is at 2.8 with a degree of freedom (d.f.) at 2, and a p value = 0.24. Therefore, it can be comprehended that the items in the *Tajwed* dimension can measure ratees' abilities accurately, reliably, and classify them based on their level of difficulty.

***Third,* Rater.** Data obtained from raters is computed by Facet software and produces a summary report in Table 4.9. In general, the raters who assessed the Quranic recitation performance in the *Tajwed* dimension gave an assessment that was consistent and quite good in terms of discrimination. The RMSE value of 0.08 indicates that the estimation of the rater's ability parameters was performed with a low error rate (Linacre, 2023). The variation of the raters in providing assessments was also in an acceptable position for human-based performance assessments, because the standard deviation (SD) value was 0.21 (Engelhard & Wang, 2024). The separation ratio value of 2.61 and the strata value of 3.82 explained that the raters could be grouped into 3 to 4 different levels (Bond et al., 2021). The levels in question address the differences between the raters, including those who exhibit both sufficient severity and leniency in providing assessment scores (Linacre, 2023). Therefore, the differences between the raters in assessing can be classified and modelled in a structured manner to minimize the possibility of distortion in the

assessment's validity.

The reliability value of the rater group is quite high, 0.87, reflecting the internal consistency of the model in distinguishing performance between raters. The reliability is not-inter-rater, which means it focuses more on the consistency of parameter estimates, so it does not refer to the results of direct agreement between raters (Fisher, 2007). The results of the chi-square test of the fixed model 33.3 with a degree of freedom of 3 showed a significant difference, $p = 0.00$, indicating that the assessment methods of the raters varied (Linacre, 2023). In addition, the inter-rater agreement was found to be 37.6% of the 1.200 assessment opportunities, with a data expectation of 34.0%. The actual agreement, which turned out to be higher than expected by statistical measurements, showed that there was still significant disagreement between raters in the assessment, although the difference in this report was still relatively low (Linacre, 2023). This report indicates the need for assessment training or the use of more standardized and explicit rubrics. The objective is to enable the raters to calibrate their perception of the assessment (Engelhard & Wang, 2024). Hence, the summary report to the raters demonstrates that, despite there being good consistency in the model and discrimination between raters, policy intervention in the form of training or explicit standardization of the assessment is imperative. This necessity arises as a fundamental component in ensuring the principles of assessment, which are valid, reliable, and fair within the framework of performance-based evaluation.

Fashahah dimension

Table 4. 10 *Ratee, item, and rater measurement report in the Fashahah dimension*

	Ratee*	**Item**	**Rater**
N	50	5	4
Root Mean Square Error (RMSE)	0.65	0.31	0.12
Adjusted (True) Standard Deviation (SD)	0.68	1.27	0.31
Separation Ratio	1.04	4.12	2.62
Strata	1.72	5.83	3.83
Reliability Index	0.52	0.94	0.87
Model fixed chi-square	104.1	47.0	25.9
Significance (Probability) (p)	0.00	0.00	0.00
Degree of freedom (d.f.)	49	3	3
Model random effect chi-square	27.9	2.7	2.7

Significance (Probability) (p)	0.99	0.25	0.26
Degree of freedom (d.f.)	48	2	2
Inter-Rater agreement opportunities	-	-	1108
Observed Exact Agreements (%)	-	-	70.2%
Expected Agreements (%)	-	-	70.7%

*The reports include the data extremes

Previously, it was revealed that the dataset on the *Fashahah* dimension was estimated and provided good results (see sub-chapter 4.1.2). However, improvements in several aspects must be made to suppress residual measurements. For this reason, analysing the statistical report in Table 4.10 on the facets level must be conducted to see the consistency and precision of measurement at the detail level. ***First, Ratee.*** Referring to Table 4.10, the Root Mean Square Error (RMSE) value shows that the level of measurement error is at 0.65. This value indicates a good accuracy of the model estimation to the data (Linacre, 2023). The adjusted (true) standard deviation (SD) value of 0.68 indicates that the variability in the abilities of the measured ratees is not too large. The small difference in the RMSE and SD values indicates that the variation in ratee scores may be due to measurement errors rather than true differences in participant abilities (Linacre, 2023). This important finding suggests that the ratee's ability in Quranic recitation, particularly in the *Fashahah* aspect, is not accurately depicted, possibly due to the quality of the rubric or the assessment process carried out by the raters.

The conditions are further emphasized by the relatively low separation value of 1.04 and the strata of 1.72. This indicates that the model can only discriminate ratees' abilities into two groups. This value is relatively low because it is below the recommended standard of separation ≥ 2.0 and strata ≥ 3.0 (Bond et al., 2021). In other words, many participants are considered to have similar abilities, even though they may not in reality. This is problematic, especially in the context of competition that requires fair and accurate assessments. Not only that, but the reliability value is also only 0.52, which is low (Fisher, 2007). This value represents the consistency of the participant's ability estimates; therefore, this reliability value is not strong enough to be used for individual decision making, such as champion rankings. However, the results of the Fixed chi-square model, which is at a value of 25.9 with 3 degrees of freedom, and significance of $p = 0.00$, are significant, indicating real variation in ratee performance. Therefore, further analysis needs to be conducted, given that the results of this data lead to consequential validity and reliability that question whether the decision taken in assessing the ratee is fair and accurate.

Second, Item. Based on the statistical results shown in Table 4.10, the items in the *Fashahah* dimension are highly reliable with a value of 0.94. According to Fisher (2007), this value indicates very good consistency in measuring differences in item difficulty levels. The separation value of 4.12, according to Bond et al. (2021), shows that the items in the *Fashahah* dimension can distinguish various levels of difficulty. A strata value of 5.83 confirms that these items can group the level of difficulty with quite high granularity (Bond et al., 2021).

According to Linacre (2023), the RMSE value of 0.31 shows that the item difficulty estimate is quite stable and accurate. The fixed chi-square test result of 47.0 with 3 degrees of freedom and a *p* value of 0.00 confirms that the data used is significantly able to differentiate items with levels that vary systematically (Linacre, 2023). In contrast, the random chi-square test result of 2.7 with 2 degrees of freedom and a *p* value of 0.25 indicates that, if the assumption of the item difficulty distribution is considered random or normal, then there is no significant difference (Linacre, 2023). Referring to Linacre (2023), this condition shows that the variability of item difficulty in the data is primarily due to systematic factors rather than random fluctuation.

Third, Rater. Based on the MFRM analysis results produced by Facet software, the rater's data recorded a good level of reliability, with a value of 0.87. According to Fisher's (2007) standard, this data is considered good because it consistently distinguishes the behaviour of raters when assigning scores. The separation value of 2.62 and the strata value of 3.83 indicate that the model can differentiate rater performance (Bond et al., 2021). A strata value of 3.83 shows that the rater's performance levels in giving scores can be categorized into more than three groups. The low RMSE value of 0.12 explains that the model provides fairly accurate predictions of rater ability (Linacre, 2023). According to the explanation of Engelhard and Wang (2024), the adjusted (true) SD of 0.31 also shows that there is enough variability in the data to identify differences in ability among raters.

According to Linacre (2023), the fixed chi-square test was 25.9 with a degree of freedom of 3 and a *p* value of 0.00 and the random fixed chi-square test value was 2.7 with a degree of freedom of 2 and a *p* value of 0.26 indicate that the differences between raters are not occur randomly, but rather due to a real or systematic pattern in the data. Of the 1.108 assessment opportunities, 778 rater decisions aligned with the model. The observed agreement was 70.2% compared to the expected agreement of 70.7%. Referring to Engelhard and Wang (2024), these results suggest that raters are quite consistent in

providing assessments because the difference between observed agreement and expected agreement percentage is quite small. This suggests that raters agree with each other to a high degree and do not exhibit extreme bias when giving scores to ratees.

Lagu dimension

Table 4. 11 *Ratee, item, and rater measurement report in Lagu dimension*

	Ratee	**Item**	**Rater**
N	50	5	4
Root Mean Square Error (RMSE)	0.44	0.16	0.11
Adjusted (True) Standard Deviation (SD)	1.21	1.61	0.27
Separation Ratio	2.73	10.03	2.54
Strata	3.97	13.70	3.72
Reliability Index	0.88	0.99	0.87
Model fixed chi-square	458.9	536.4	32.8
Significance (Probability) (p)	0.00	0.00	0.00
Degree of freedom (d.f.)	49	4	3
Model random effect chi-square	42.4	4.0	2.8
Significance (Probability) (p)	0.70	0.26	0.25
Degree of freedom (d.f.)	48	3	2
Inter-Rater agreement opportunities	-	-	1500
Observed Exact Agreements (%)	-	-	61.9%
Expected Agreements (%)	-	-	58.6%

In this section, the statistical report at the facets level on the *Lagu* dimension is reported in Table 4.10. ***First, Ratee***. Referring to Table 4.11 in the reporting section of ratee statistics, the reliability value is quite high at 0.88. According to Fisher (2007), this shows that the measurement system effectively distinguishes between levels of ratee ability. The separation value of 2.73 and the strata value of 3.97 indicate that the ratee's ability can be grouped into three different categories. According to Bond et al. (2021), this figure is sufficient to group ratees into different strata.

From an RMSE perspective, Linacre (2023) suggests that a model with a value of 0.44 is considered to have a very good level of estimation. The adjusted (true) standard deviation value also explains that the data has sufficient variability to distinguish the

abilities of the ratees because it has a value of 1.21. Referring to the explanations of Linacre (2023) and Engelhard and Wang (2024), these two results demonstrate that the ratee's ability estimates are quite accurate and stable, enabling the model to predict the expected results. Additionally, the variability index provided by the data is not too deviant to impact the imbalance of variation in the model.

The other report was the fixed chi-square test result with a value of 458.9, a degree of freedom of 49, and a *p*-value of 0.00. This index shows that the difference in ratee ability is significant, indicating that the variation between ratees is real. The random chi-square emphasizes this, with a value of 42.4 with a degree of freedom of 48 and a *p* value of 0.70. In other words, the variability of ratee ability does not come from a random distribution.

Second, Item. The analysis conducted by MFRM shows that the modelled data has a very high reliability, which is at 0.99. According to Fisher (2007), this high number indicates that the measurement system consistently distinguishes items according to their level of difficulty. The separation value of 10.03 indicates that the separation of the item difficulty levels is excellent. Additionally, the strata value of 13.70 explains the extraordinary capacity of items in grouping item difficulty levels into up to more than 13 groups and even to more granular levels (Bond et al., 2021).

From an RMSE perspective, the value of 0.16 owned by these items is quite low. According to Linacre (2023), this figure indicates that the difficulty items in this dimension can be accurately estimated by the model. The adjusted (True) SD value of 1.61 also confirms good variation in the data, indicating a wide distribution of item difficulty levels in the dataset. Meanwhile, the fixed chi-square value of 536.4 with a degree of freedom of 4 and a *p* value of 0.00 makes it clear that there is a significant difference in how item difficulty is measured. The random chi-square result of 4.0, with a degree of freedom of 3 and a *p* value of 0.26, shows that the distribution of items in the data is a systematic pattern captured by the model, rather than the result of random factors. These results suggest that the model is stable.

Third, Rater. Referring to Table 4.11 of the rater statistics report, the dataset related to the rater shows a good level of reliability in evaluating consistency between raters. With a value of 0.87, it demonstrates good reliability quality (Fisher, 2007). The separation value is at 2.54, and the strata value is at 3.72. According to Bond et al. (2021), these two values show that the model can distinguish the level of severity between raters and group raters into three groups with different assessment patterns.

The resulting RMSE value is 0.11. Referring to Linacre's explanation (2023), these results indicate that the error in estimating the rater's severity is very small. Thus, it can be understood that in distinguishing the rater's severity characteristics has been done with a good level of precision has been achieved. The adjusted (True) SD value of 0.27 explains that, although there are differences in the raters' assessment patterns, the variability is still relatively controlled. The results of the fixed chi-square test of 32.8 with 3 degrees of freedom and a *p* value of 0.00 also indicate that the difference in the rater assessment patterns is significant. The random chi-square test results of 2.8 with 2 degrees of freedom and a *p* value of 0.25 also explain that the difference in rater assessment pattern stems from a systematic difference in how raters give scores. Of the 1.500 assessment opportunities, the inter-rater agreement report produced a level of agreement of 61.9%. This result is higher than the expected agreement by the model of 58.6%. The difference is not significant, indicating that the pattern is quite stable, though the percentage of agreement between raters could be maximized. The measurement results show that the model can estimate that the responses in the *Lagu* dimension are very stable and reliable in measuring the rater assessment pattern.

Suara dimension

Table 4. 12 *Ratee, item, and rater measurement report in the Suara dimension*

	Ratee	**Item**	**Rater**
N	50	5	4
Root Mean Square Error (RMSE)	0.54	0.13	0.11
Adjusted (True) Standard Deviation (SD)	1.11	0.92	0.38
Separation Ratio	2.07	7.03	3.31
Strata	3.10	9.70	4.75
Reliability Index	0.81	0.98	0.92
Model fixed chi-square	176.6	262.8	46.5
Significance (Probability) (p)	0.00	0.00	0.00
Degree of freedom (d.f.)	49	4	3
Model random (normal) chi-square	33.5	3.9	2.8
Significance (Probability) (p)	0.94	0.27	0.24
Degree of freedom (d.f.)	48	3	2
Inter-Rater agreement opportunities	-	-	1500
Observed Exact Agreements (%)	-	-	56.3%

Expected Agreements (%)	-	-	55.0%

Further analysis is given to the group of facets on the *Suara* dimension reported in Table 4.12. ***First, Ratee***. Refer to Table 4.12, the measurement estimation carried out by the model produces a fairly precise value because the RMSE is relatively low at 0.54. The adjusted (true) standard deviation is 1.11. The separation index and strata values of 2.07 and 3.10, respectively, indicate that the model can differentiate individual abilities into three levels of difference with a fairly good level of precision (Bond et al., 2021). The reliability value is 0.81, which, according to Fisher (2007), indicates that the data consistently measures ratee abilities. The chi-square analysis for the model with fixed assumptions yielded a value of 176.6 with 49 degrees of freedom and a significance of p = 0.00. These results show that there is a significant difference between the observed data and the model used. In contrast, the chi-square analysis for the model with random assumptions has a value of 33.5 with 48 degrees of freedom, and a significance of p = 0.94. These results suggest that there is no significant difference between the model assuming a normal distribution and the observed data. The measurement report results for the ratee indicate that the modelled data are good in terms of consistency and precision.

Second, Item. Based on the item measurement report data, the items used have excellent psychometric properties. The RMSE value of 0.13 is fairly low, indicating a low level of measurement error. This suggests that parameter estimation is quite accurate (Linacre, 2023). The variability in the level of item difficulty is also shown in the adjusted (true) standard deviation value of 0.92, which indicates that the distribution of items in the measurement is good. Not only that, the separation and strata values of the items in the *Suara* dimension have a high ability to distinguish levels of difficulty at 7.03 and 9.70, respectively (Linacre, 2023). These results demonstrate that the model can distinguish more than nine different levels of item difficulty. The reliability value is also reported as a very good figure, 0.98. According to Fisher (2007), this figure confirms that the items measured by the model have strong internal consistency. The chi-square test for the fixed model has a value of 262.8 with 4 degrees of freedom, and a significance of p 0.00. This result indicates that the assumption of the same level of item difficulty is rejected because, in actual measurements, there is real variation between items. Conversely, the chi-square test results on the random model show that the data are not significantly different from the model when assuming a normal distribution of item difficulty. According to Linacre (2023), these results confirm that the model corresponds to the empirical data collected.

***Third,* Rater**. Table 4.12 provides information on the quality and consistency of the assessment between raters. The RMSE value of 0.11 indicates that the measurement error is very small, suggesting that a precise measurement estimate (Linacre, 2023). The variation among raters in terms of severity levels shows real variation with an adjusted (true) standard deviation value of 0.38. The separation and strata values are at 3.31 and 4.75, respectively, confirming that the rater's severity level can be distinguished into four different groups (Bond et al., 2021). The reliability value is also high at 0.92, indicating very good consistency in the measurement between raters (Fisher, 2007).

The chi-square test for the fixed model has significant results, a chi-square value of 46.5 with a degree of freedom of 3 and a significance of $p = 0.00$. This explains that the assumption of rater severity similarity is rejected, so that the data show differences in how raters assess. The random model test results were not significant, chi-square 2.8, degrees of freedom 2, and significance probability $p = 0.24$. These findings indicate that the variation between raters is not substantial or statistically significant, indicating that differences between raters occur randomly and do not systematically influence the assessment (Linacre, 2023; Eckes, 2015).

One significant report is the inter-rater agreement, which provided informative results. Of the 1.500 opportunities for agreement, 845 or 56.3% were recorded as the percentage of observed agreements. Meanwhile, the expected agreement by the model is 55.0%. Referring to Linacre's explanation (2023), the size of the difference determines the consistency with which a rater gives an assessment. In this case, the difference between observed and expected agreement is 1.3%. This indicates that the assessment pattern is in line with the model, meaning the raters understand the scale and provide consistent scores (Engelhard & Wang, 2024; Linacre, 2023). However, the results of the random chi-square test must be noted.

Overall, the statistical indicator reports at the facet level across all dimensions yielded very good reliability results. Most results report that the data explaining the ratee's performance is reliable and well-separated, except for the ratees in the *Fashahah* dimension, which requires further investigation due to its low value. On the other hand, the reliability of the items shows excellent results. Even in the *Lagu* and *Suara* dimensions, the consistency and the precision are very good because the reliability and separation values are very high. The reliability results given on the rater group are also good, even though the assessment pattern and inter-rater agreement are varied. To answer the research

question, the Quranic recitation assessment rubric has a very good reliability value, indicating that the rubric is reliable and consistent across various measurement conditions provided by raters from diverse backgrounds when assessing various participants.

4.2.3. Fairness of the Quranic Recitation Assessment

The next psychometric concept used to evaluate the quality of a rubric is fairness. Referring to The Standards (2014) on which this research analysis is based, the fairness aspect targets the rubric's ability to make fair assessments. In other words, meaning does not differ based on individual backgrounds. This is important to reduce unnecessary variation and encourage accurate score interpretations for the intended purposes (The Standards, 2014). Therefore, the third question is posed to reveal this aspect: *To what extent does the assessment ensure fairness for all participants, considering judge severity, item difficulty, and unexpected scoring patterns revealed through MFRM analysis?*

To answer the last question, evidence will be collected in several stages. The analysis begins by reporting the model-data fit of the rater and ratee. Then, the analysis continues by identifying the interaction between the rater and ratee to determine if there is an indication of a differential item or person functioning or a pattern of bias. The final stage will be an exploration of unexpected responses reported by the software. All these stages are important because the rubric will be considered fair if it does not provide room for bias in the assessment. The Standards (2014) mentioned that the content, context, and response in assessment can be the source of evidence but not limited to. For this reason, the findings of the item fit in RQ1 will also be selectively used as evidence to interpret the relationship between items, raters, and ratees that can affect the fairness in the assessment process.

Evidence in the rater and ratee fit analysis

Table 4. 13 *Rater fit measurement report across dimensions*

Rater	Total score	Model		Mean square		Z standard.		Cor.	Exact agreement		
		Mea	S. E	In	Out	In	Out	PtMea	Obs.	Exp.	Dist
Tajwed											

SD	685	-0.29	0.07	0.87	0.84	-1.3	-1.6	0.57	35.2%	32.3%	2.9
NA	756	-0.71	0.08	1.28	1.19	2.4	1.6	0.50	41.3%	34.6%	6.7
AY	752	-0.69	0.08	0.70	0.71	-3.1	-2.9	0.59	37.0%	34.6%	2.4
MA	785	-0.91	0.08	1.20	1.18	1.7	1.5	0.41	36.8%	34.7%	2.1
Mean	744.5	-0.65	0.08	**1.01**	**0.98**	**-0.1**	**-0.4**	**0.52**			
Fashahah											
SW	698	-1.82	0.10	1.22	0.95	1.4	0.0	0.45	66.9%	69.1%	-2.2
QN	727	-2.12	0.11	1.05	0.98	0.3	0.0	0.36	70.6%	71.0%	-0.4
RI	611	-1.74	0.11	0.82	0.50	-1.0	-1.4	0.45	70.3%	70.3%	0.0
YN	743	-2.58	0.15	0.97	0.44	0.0	-1.1	0.35	73.0%	72.4%	0.6
Mean	694.8	-2.07	0.12	**1.01**	**0.72**	**0.2**	**-0.6**	**0.40**			
Lagu											
NI	1096	-2.94	0.11	1.06	0.73	0.5	-1.0	0.70	62.9%	59.4%	3.5
PB	1098	-2.96	0.11	1.10	1.19	0.8	0.7	0.69	61.5%	59.5%	2.0
AM	1032	-2.24	0.10	1.39	0.95	3.2	-0.1	0.74	60.4%	56.1%	4.3
AP	1086	-2.82	0.11	0.86	0.64	-1.2	-1.5	0.74	62.9%	59.2%	3.7
Mean	1078	-2.74	0.11	**1.10**	**0.88**	**0.8**	**-0.5**	**0.72**			
Suara											
UM	645	-1.74	0.12	0.75	0.73	-2.7	-1.3	0.61	59.1%	56.0%	3.1
MF	570	-0.78	0.11	1.26	1.44	2.8	2.8	0.49	51.1%	53.5%	-2.4
JM	621	-1.40	0.12	0.89	0.80	-1.1	-1.1	0.60	60.7%	56.1%	4.6
WM	579	-0.88	0.11	1.02	0.92	0.2	-0.4	0.62	54.5%	54.3%	0.2
Mean	603.8	-1.20	0.11	**0.98**	**0.97**	**-0.2**	**0.97**	**0.58**			

Overall, Table 4.13 shows that most raters provide assessment patterns that are accepted by the model. Ten of 16 raters were recorded as having MnSq, ZSTD, and point measure correlation values that were within the acceptable range (Engelhard & Wind, 2018; Linacre, 2002; Boone et al., 2015). According to Boone et al. (2015), if only the MnSq and point measure correlation values (note: if the sample size is big), 15 raters were recorded as having good data fit. Only rater YN was detected as overfit because the outfit MnSq value was below the standard (0.44). This means that rater YN gave an assessment that was too predictable or too consistent, so that it was unable to provide new information in the assessment. However, if highlighting the ZSTD value as well, there were seven raters, including YN, who were recorded as having values outside the ideal range. For raters NA, AM, and MF, it was recorded that they had a higher value than the ideal range, which was more than +.2.0 logit, which, according to Linacre (2002), the data was detected as unpredictable or inconsistent. This result can be an indication that the raters may be too lenient or severe in providing several assessment patterns. Meanwhile, AY and UM were recorded as having ZSTD values both in terms of infit and outfit that were lower than the ideal standard (< - 2.0). According to Linacre (2002), the measurement patterns given by these two raters were considered too consistent, so that other aspects may be constraining the response pattern, because this condition can indicate systematic bias.

MFRM analysis of the rater response patterns also provides a report on the observed and expected agreement for each rater. In the *Tajwed* dimension, rater NA was recorded as having a pattern that turned out to be much higher than the model's expectations. While other raters tended to have similar distances, comparable conditions appeared across other dimensions, indicating that the observed rater assessment patterns did not significantly differ from the model's expectations. Meanwhile, in the *Fashahah* dimension, the exact agreement distance was quite small, indicating that bias may be very minimal in this dimension.

Further analysis was given to 50 ratees in Table 4.14. The analysis was carried out to provide a comprehensive picture of the data fit conditions with the model across all dimensions. The statistical values of MnSq, ZSTD, and point measure correlation guided this analysis. Overall, the response patterns owned by the ratees were quite diverse but still acceptable if referring to the final mean results in each dimension. There were three ratees, numbers 5, 6, and 21, who showed extreme patterns in the *Fashahah* dimension. Even for ratee number 21, an extreme pattern was also detected in the *Suara* dimension. Referring to the Wright map report in Figures 4.12 and 4.14, these ratees were indeed in a very high

logit position compared to the others. The extreme pattern indicates that these three raters performed too well, causing the model to identify them as outliers.

Regarding the MnSq indicator, all assessment patterns given to the ratee indicate that only five ratees, namely, 7, 13, 17, 24, and 38, show patterns predicted by the model. The rest provide various patterns, either too predictable (overfit) or unpredictable (misfit) by the model. Meanwhile, in contrast to the ZSTD indicator, the ratees recorded as meeting the standard are more, 35 ratees. They are considered to have patterns that have reasonable predictability (Linacre, 2002). The other 15 raters are recorded as providing overfit and misfit measurements in several dimensions. Regarding the point measure correlation indicator, very good results are given because only 5 ratees in the *Tajwed* dimension are recorded as having patterns that are not in line with the construct expected.

The results explain that even though there is a pattern of assessment of the ratee's performance that is recorded as overfit or misfit, especially in the MnSq indicator, the model can still accept the data pattern because, if referring to the mean value of each indicator in each dimension, the range of values given is still within acceptable ranges. Some of them show very close to ideal values, approaching 1.00 for MnSq and 0.00 for ZSTD. However, apart from that, the rater and ratee fit analysis results recommend further analysis of the interaction between raters and ratees on assessment items. This is necessary to clarify whether the condition of data fit affects the fairness of the rubric, particularly in creating bias in the assessment.

Table 4. 14 *Ratee fit measurement report across dimensions*

No	Mean square								Z-standardized								Correlation			
	Infit				Outfit				Infit				Outfit				Point measure			
	T	F	L	S	T	F	L	S	T	F	L	S	T	F	L	S	T	F	L	S
1	1.00	0.59	0.91	1.45	0.92	0.33	0.47	1.73	0.1	-0.2	0.0	1.1	0.0	0.2	0.0	1.2	0.34	0.27	0.41	0.37
2	0.69	1.08	0.47	1.32	0.73	0.77	0.51	1.13	-0.6	0.3	-1.3	1.0	-0.6	0.3	-0.3	0.4	0.63	0.36	0.64	0.53
3	0.60	1.36	0.41	1.14	0.62	0.68	0.36	1.33	-1.1	0.7	-1.8	0.5	-1.0	0.0	-1.4	1.1	0.43	0.40	0.83	0.44
4	0.42	0.79	0.69	0.87	0.40	1.47	0.55	1.09	-1.6	0.0	-0.7	-0.3	-1.8	0.8	-0.6	0.3	-0.22	0.20	0.71	0.53
5	1.79	Max	0.92	0.48	1.73	Max	0.49	0.25	1.7	Max	0.1	-0.9	1.6	Max	0.3	-0.9	0.18	0.00	0.34	0.67
6	1.14	Max	0.96	1.07	1.11	Max	0.50	1.43	0.4	Max	0.1	0.4	0.4	Max	0.3	0.7	0.26	0.00	0.33	0.02
7	0.67	1.52	1.80	1.24	0.65	1.04	1.04	1.27	-0.7	0.8	1.7	0.8	-0.8	0.5	0.2	0.8	0.07	0.25	0.61	0.40
8	1.69	1.06	0.56	0.76	1.79	1.41	0.37	0.74	1.6	0.3	-1.1	-0.8	1.7	0.7	-1.0	-0.8	-0.32	0.28	0.76	0.72
9	1.30	0.98	1.55	0.85	1.49	0.58	1.17	0.78	0.8	0.2	1.3	-0.4	1.2	0.0	0.4	-0.7	0.18	0.34	0.64	0.80
10	1.74	1.95	1.54	1.25	1.75	1.81	1.26	1.12	1.9	1.7	1.3	0.9	1.9	1.0	0.6	0.4	0.62	0.29	0.67	0.67
11	0.33	0.66	0.93	0.31	0.36	0.60	0.49	0.29	-2.1	-0.3	0.0	-2.5	-2.0	0.0	-0.2	-1.7	0.05	0.42	0.55	0.82
12	0.51	0.79	0.92	0.76	0.55	0.33	0.56	0.24	-1.3	0.2	0.0	0.0	-1.2	0.5	-0.2	-0.2	0.22	0.19	0.57	0.42
13	0.60	1.07	1.04	1.17	0.59	0.82	0.72	1.16	-1.0	0.3	0.2	0.6	-1.0	0.1	-0.5	0.5	0.26	0.35	0.76	0.52
14	0.71	0.48	0.42	1.03	0.63	0.29	0.35	1.08	-0.7	-0.5	-1.6	0.1	-0.9	0.0	-1.2	0.3	0.57	0.42	0.80	0.22
15	0.93	0.95	0.83	0.92	0.86	0.79	0.48	1.23	0.0	0.1	-0.2	-0.1	-0.2	0.2	-0.6	0.6	0.23	0.32	0.69	0.38
16	1.80	0.18	0.96	1.00	1.73	0.14	0.45	0.89	1.6	-1.6	0.2	0.2	1.6	-0.6	0.6	0.2	0.29	0.55	0.25	0.22

17	0.96	0.58	0.88	0.93	0.96	0.78	0.73	0.85	0.0	-0.9	-0.2	-0.1	0.0	0.0	-0.6	-0.4	0.55	0.43	0.81	0.79
18	1.58	0.89	1.31	0.95	1.53	0.59	0.75	0.50	1.3	0.3	0.7	0.2	1.2	0.7	0.1	0.1	0.26	0.11	0.45	0.26
19	1.28	0.26	0.54	1.44	1.24	0.23	0.34	1.42	0.8	-2.0	-1.2	1.3	0.7	-0.9	-1.2	1.1	0.57	0.61	0.78	0.35
20	0.73	0.36	1.49	1.74	0.66	0.27	0.81	1.21	-0.5	-1.3	1.1	1.6	-0.8	-0.6	0.0	0.5	-0.3	0.51	0.60	0.31
21	0.69	Max	0.53	Max	0.76	Max	0.31	Max	-0.6	Max	-1.1	Max	-0.5	Max	-0.6	Max	0.18	0.00	0.67	0.00
22	1.37	1.18	1.56	0.59	1.34	0.67	0.86	.047	1.0	0.4	1.3	-1.3	0.9	0.2	0.0	-1.3	0.48	0.27	0.61	0.72
23	0.80	0.94	0.58	0.63	0.76	0.42	0.39	0.61	-0.3	0.1	-1.0	-1.3	-0.5	0.0	-1.1	-1.1	0.51	0.34	0.79	0.78
24	1.02	0.92	1.27	0.62	1.04	0.75	0.65	0.53	0.1	0.3	0.7	-1.3	0.2	0.8	-0.1	-1.4	0.42	0.21	0.54	0.87
25	0.86	0.38	0.28	1.36	0.85	0.44	0.27	1.49	-0.2	-1.2	-2.6	1.2	-0.2	-0.3	-2.0	1.4	0.17	0.50	0.89	0.43
26	0.36	2.11	0.94	1.09	0.31	1.13	0.71	1.08	-2.1	2.1	0.0	0.4	-2.4	0.4	-0.7	0.3	0.42	0.41	0.82	0.59
27	0.72	2.57	1.12	1.43	0.64	1.30	0.94	1.47	-0.8	2.0	0.4	1.5	-1.0	0.6	0.0	1.5	0.38	0.35	0.81	0.46
28	0.85	0.90	0.47	0.92	0.90	0.97	0.35	1.00	-0.3	0.0	-1.5	-0.1	-0.1	0.4	-1.4	0.1	0.33	0.30	0.81	0.59
29	1.16	0.62	0.78	0.96	1.23	0.40	0.70	1.31	0.5	-0.1	-0.4	0.0	0.7	0.2	-0.1	0.7	0.26	0.35	0.59	0.38
30	0.59	0.89	0.59	1.10	0.66	0.65	0.42	1.03	-1.0	0.1	-0.9	0.3	-0.8	0.3	-0.8	0.2	0.55	0.29	0.73	0.41
31	1.03	0.41	1.01	0.84	1.04	0.24	0.96	1.15	0.2	-1.0	0.1	-0.1	0.2	-0.6	0.3	0.4	0.53	0.47	0.36	0.31
32	0.49	0.79	0.99	0.73	0.49	0.33	0.54	0.33	-1.3	0.2	0.2	-0.2	-1.4	0.5	0.3	-0.4	0.79	0.19	0.31	0.49
33	1.08	1.25	0.74	0.92	1.07	0.39	0.38	0.90	0.3	0.5	-0.4	0.0	0.3	0.2	-0.5	0.0	0.02	0.27	0.62	0.35
34	0.76	0.62	1.20	0.69	0.78	0.40	0.77	0.67	-0.5	-0.1	0.5	-1.0	-0.4	0.2	-0.1	-0.9	0.56	0.35	0.64	0.79
35	1.01	0.65	1.66	0.90	1.04	0.31	0.57	0.96	0.1	-0.2	-0.1	-0.2	0.2	-0.1	-0.7	0.0	0.16	0.41	0.71	0.53
36	0.69	0.63	0.90	1.32	0.68	0.40	2.74	1.34	-1.0	-0.6	3.4	1.1	-1.0	-0.4	3.4	1.1	0.37	0.49	0.50	0.46

37	0.60	1.07	2.60	1.11	0.61	0.80	1.58	1.08	-1.6	0.3	1.8	0.4	-1.5	0.0	1.6	0.3	0.75	0.48	0.72	0.43
38	0.95	1.05	0.68	0.71	1.00	0.82	0.60	0.73	0.0	0.2	-0.8	-1.0	0.1	0.0	-0.7	-0.8	0.19	0.43	0.76	0.68
39	0.96	2.06	2.70	1.04	0.87	3.53	4.50	1.03	0.0	2.1	3.8	0.2	-0.2	2.2	6.0	0.1	0.59	0.06	0.14	0.61
40	0.91	0.96	0.62	1.36	0.89	0.63	0.61	1.65	-0.2	0.0	-1.0	1.2	-0.2	-0.1	-0.8	1.8	0.49	0.49	0.78	0.25
41	0.89	0.48	0.34	1.13	0.90	0.29	0.32	1.03	-0.1	-0.5	-2.1	0.4	-0.1	0.0	-1.5	0.2	0.44	0.42	0.85	0.39
42	1.55	0.51	0.30	1.05	1.64	0.29	0.22	1.08	1.2	-0.4	-2.2	0.2	1.4	0.0	-1.5	0.3	-0.15	0.41	0.83	0.43
43	1.47	1.02	0.48	0.81	1.39	0.58	0.38	0.83	1.1	0.2	-1.4	-0.6	1.0	0.0	-1.2	-0.4	0.16	0.38	0.79	0.69
44	1.57	1.66	2.56	1.09	1.57	1.02	2.77	1.14	1.7	1.2	3.6	0.4	1.6	0.3	3.8	0.5	0.13	0.39	0.37	0.25
45	1.68	0.68	1.34	1.02	1.64	0.66	1.52	1.09	2.0	-0.4	1.0	0.1	1.8	0.0	1.5	0.3	0.32	0.45	0.65	0.45
46	0.87	0.76	1.13	0.87	0.93	0.43	1.02	0.92	-0.3	-0.4	0.4	-0.3	-0.1	-0.5	0.1	-0.1	-0.11	0.57	0.78	0.47
47	0.91	0.74	1.17	0.94	0.98	0.45	1.19	0.88	-0.1	-0.3	0.5	-0.1	0.0	-0.3	0.6	-0.2	0.12	0.48	0.71	0.63
48	0.97	0.45	1.03	0.63	0.87	0.27	1.09	0.60	0.0	-0.7	0.2	-1.3	-0.2	-0.3	0.3	-1.1	0.21	0.46	0.56	0.73
49	0.67	1.14	0.43	1.08	0.63	0.72	0.74	0.97	-1.0	0.4	-1.6	0.3	-1.1	0.0	-0.2	0.0	0.22	0.48	0.73	0.72
50	1.14	1.56	1.61	0.52	1.14	3.45	3.38	0.54	0.5	1.5	1.5	-2.0	0.5	2.6	3.7	-1.8	0.38	0.04	0.49	0.76
M	**0.98**	**0.95**	**1.01**	**0.98**	**0.98**	**0.76**	**0.88**	**0.97**	**0.0**	**0.1**	**0.0**	**0.0**	**-0.1**	**0.2**	**0.0**	**0.1**	**0.30**	**0.34**	**0.64**	**0.50**

*T: *Tajwed*, F: *Fashahah*, L: *Lagu*, S: *Suara*
*The underline: Overfit or misfit

Evidence from the bias analysis

Table 4. 15 *Bias/interaction report across dimensions*

Rater	Ratee	Obs. score	Exp. score	Bias size	S.E.	t	d.f.	Prob.	Mean square	
									Infit	Outfit
Tajwed: 2 out of 6										
SD	39	9	13.78	-1.03	0.46	-2.22	3	0.1127	0.5	0.5
MA	5	16	18.16	-1.11	0.62	-1.79	3	0.1707	1.7	1.6
Fashahah does not have bias										
Lagu: 4 out of 26										
AP	36	9	13.21	-1.64	0.67	-2.44	4	0.0709	1.2	3.1
AP	39	22	17.98	1.67	0.75	2.22	4	0.0907	0.3	0.3
NI	36	19	13.55	1.84	0.60	3.07	4	0.0375	1.0	0.9
PB	39	15	18.40	-1.14	0.58	-1.96	4	0.1213	6.9	8.3
Suara: 5 out of 21										
MF	3	14	9.62	2.68	1.08	2.47	4	0.0686	0.9	0.6
MF	44	13	9.62	1.80	0.84	2.15	4	0.0978	0.8	1.6
MF	42	14	11.23	1.90	1.08	1.76	4	0.1540	1.3	1.8
MF	6	14	14.65	-1.15	1.08	-1.06	4	0.3492	1.3	1.8
JM	27	13	10.68	1.29	0.84	1.55	4	0.1970	3.3	6.9

A bias analysis was performed to examine the interaction between rater and ratee, whether the rater exhibits bias towards the ratee. Table 4.15 shows several interactions that were detected as biased by considering the t value and its probability value. In addition, the mean square value is also reported to provide additional information on which raters show different patterns regardless of their fit conditions, as previously reported. Through observed and expected score information, this study can understand how big the gap is that causes bias in the assessment. Additionally, graphical design is also provided to show visual information on the bias that occurs in each dimension.

Of the total 3.760 interactions that occurred, there were 53 interactions or 1.4% indicated bias spread across all dimensions except the interactions in the *Fashahah* dimension. The *Fashahah* dimension may not show bias for several reasons, one of which is the small difference in exact agreement between the rater and the model, which narrows

the opportunity for bias to occur (see Table 4.13). The separation index can also be a factor because the indices of ratees were very small, 1.04 (see Table 4.10). Apart from that, the *Lagu* and *Suara* dimensions were recorded to provide many bias patterns as many as 26 and 21, respectively, while the *Tajwed* dimension only recorded 6 bias patterns. The statistical indicator information displayed in Table 4.15 does not report 53 bias interactions in total. The reason is based on the t value, which is still within the normal range (-2.0 to +2.0), the probability value is above 0.05, and the mean square value is not misfit (Engelhard, 2013; Linacre, 2023; Boone et al., 2014). According to Linacre (2023), this condition explains that the differences in interactions are not consistent enough to be considered bias, so the 42 out of 53 bias interactions are considered bias that are statistically not significant.

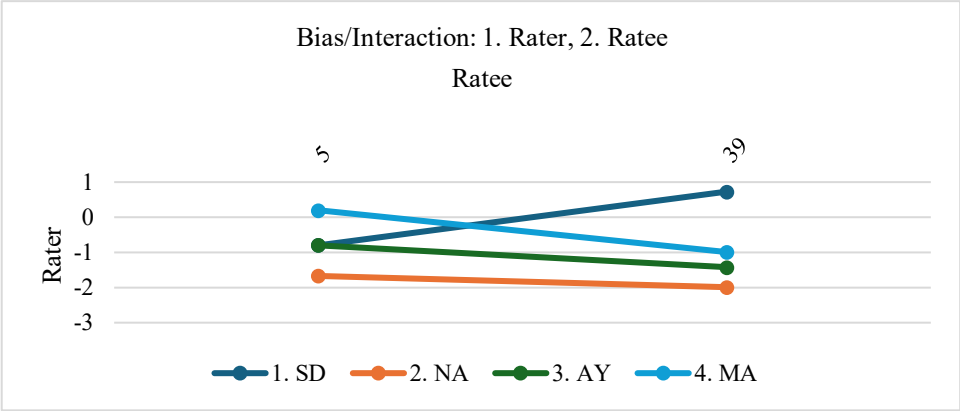
Overall, six raters' interactions were recorded to have *t*-values that were out of the standard range. This is clarified by the reports on the observed and expected scores given. For example, rater SD gave a score of 9 to ratee number 39, who, according to the model, should have given a score of 13.78; rater AP gave a score of 9 to ratee number 36, which was much lower than it should have been, 13.21. The bias pattern carried out by rater SD towards ratee number 39 and AP towards ratee number 36 is an example of bias caused by the rater being too severe in giving an assessment (Myford & Wolfe, 2004). In contrast to that, the interaction patterns between rater AP and ratee number 39, rater NI and ratee number 36, and rater MF and ratees numbers 3 and 44, show a bias caused by leniency in assessment (Myford & Wolfe, 2004). However, according to the model, this bias does not have a significant effect, except for the interaction on the *Lagu* dimension between rater NI and ratee number 36, which the model detected as having a significant bias ($p < 0.05$). The visualization of several bias patterns can be seen in Figure 4.18.

Furthermore, this study also includes the fit statistics indicator to show whether the raters who have generally been declared fit or misfit/overfit in Table 4.13, provide a similar pattern when assessed specifically. Based on Table 4.14, raters MA, NI, PB, AP, and JM are raters who are not included as raters with misfit or overfit response patterns. However, in their opportunity to interact with several ratees, they showed a misfit pattern. MA to the ratee number 5, NI to ratee number 36, PB to ratee number 39, AP to ratee number 36, and JM to ratee number 27. In contrast, rater MF, who was considered to have a misfit pattern from the previous report (Table 4.12), is also reported to have a misfit when interacting with ratees number 44, 42, and 6, indicating that assessment training for MF needs to be carried out.

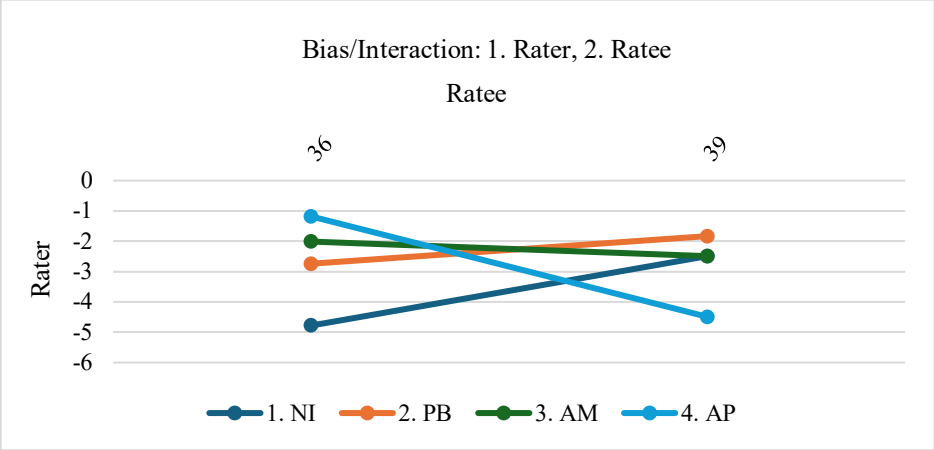
To concretely illustrate the bias that occurs, an example of interactions, NI's assessment of ratee number 36, will be clarified by the qualitative findings in the rubric (Figure 4.19). In this case, all raters gave similar comments on the ratee's inability to perform the *maqam* (song) composition. In the science of *nagham*, each type of *maqam* has a level of tone from the opening to the closing (e.g., *Maqam bayyati*: *Qarar* (basic)-*nawa* (middle)-*jawab* (high) -*jawab al-jawab* (highest)) (Salim, 2004). If a ratee is unable to perform the *maqam* according to the rules, a reduction in scores will be given. The rater's notes are as follows; NI: "*Lagu kurang jelas/ngawur* (The song is not clear/incoherent)", PB: "*Lagu kurang jelas, nawa tidak jelas, jawab bayyati-variasi kurang jelas* (The song is not clear, *nawa* tone is unclear, *jawab bayyati*-variation is unclear)", AM: "*Tidak dimulai dan diakhir dengan baik, lagu tidak bisa ditebak, suara lirih, penghayatan ngambang, variasi kosong dan polos* (Does not start and finish well, the song is unpredictable, the voice is weak, the appreciation is lacking, the variations are empty and plain)", and rater AP: "*Belum menguasai nada dan tidak jelas jenis lagu apa tersebut* (Has not mastered the tone and it is unclear what type of song it is)". From these notes, ratee number 36 is considered not to have mastered the *nagham* science in Quranic recitation according to all raters. However, while most raters assigned a total score within the range of 12 and 14, a notable deviation was observed in rater NI, who assigned a total score of 19. This indicates that rater NI was leniently biased, which according to the model is significant ($p = 0.0375$).

Further analysis of the mean square value in the bias report reveals that NI has a misfit in the MnSq infit value. According to Linacre (2023), misfit in the infit indicator is suspected to occur due to internal influences on the rater rather than the outlier factors (outfit). Qualitative data show that NI has the same understanding as the other raters, which shows that the lenient score given is not due to lucky guesses, input errors, or even confusing ratee performance patterns. According to the demographic data report, the code for NI is "L1FSulut21A6PaTilTah", while for ratee number 36 is "36FBanten12J1O". Based on this code, there is no overlapping background, emphasizing that the bias is not due to outlier factors, but rather to Ni's inconsistency or careless assessment. This study assumes that the bias occurs because the rater has repeated the assessment many times (± 7 minutes per participant), which may cause inconsistency or carelessness in the assessment. This assumption is reasonable because, based on the general rater fit statistics report, NI does not show a misfit or overfit pattern in the assessment (see Table 4.14). The report from the Wright map in the *Lagu* dimension also confirmed that the ratee number 36 position is at the bottom of the map, while NI was at the top of the ratee (Figure 4.13). In view of this explanation, it can be understood that the model of the analysis and the

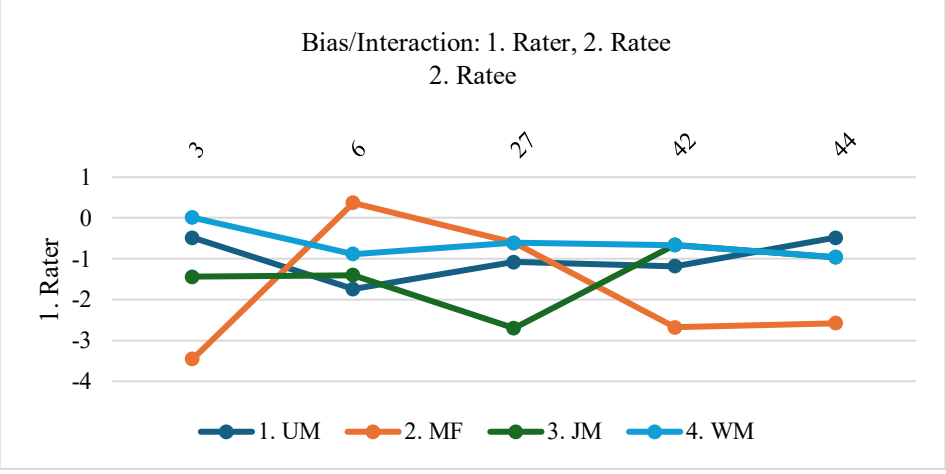
rubric can provide detailed evidence of the fair assessment.



Tajwed



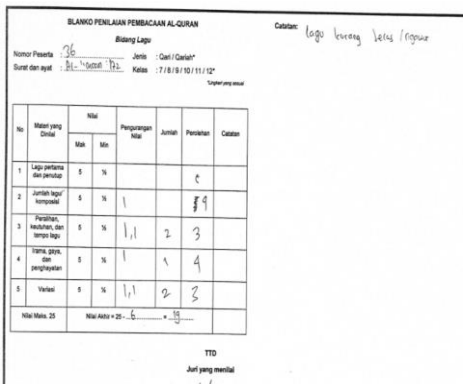
Lagu



Suara

Figure 4. 18 *Bias/interaction between rater and ratee*

Rater NI


BLANKO PENILAIAN PEMBACAAN AL-QURAN

Bidang Lagu

Nomor Peserta : 36 Jenis : Qari / Qariah*

Surat dan ayat : Al-'Alaq 1-2 Kelas : 7 / 8 / 9 / 10 / 11 / 12*

*Lingkari yang sesuai

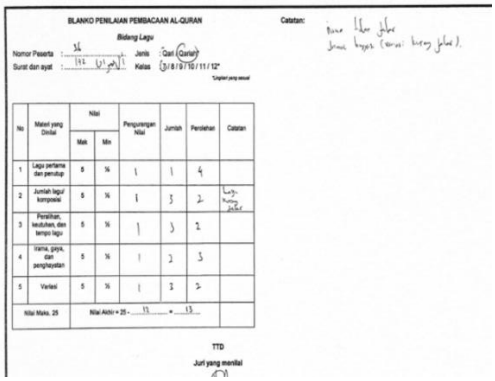
Catatan: lagu kurang lancar / tegang

No	Materi yang Dinilai	Nilai Mak	Nilai Min	Pengurangan Nilai	Jumlah	Perolehan	Catatan
1	Lagu pertama dan penutup	5	½			5	
2	Jumlah lagu/ komposisi	5	½	1		4	
3	Peralihan, keutuhan, dan tempo lagu	5	½	1,1	2	3	
4	Irama, gaya, dan penghayatan	5	½	1	1	4	
5	Variasi	5	½	1,1	2	3	
Nilai Maks. 25		Nilai Akhir = 25 - 6 = 19					

TTD

Juri yang menilai

Rater PB


BLANKO PENILAIAN PEMBACAAN AL-QURAN

Bidang Lagu

Nomor Peserta : 36 Jenis : Qari (Qariah)*

Surat dan ayat : Al-'Alaq 1-2 Kelas : 7 / 8 / 9 / 10 / 11 / 12*

*Lingkari yang sesuai

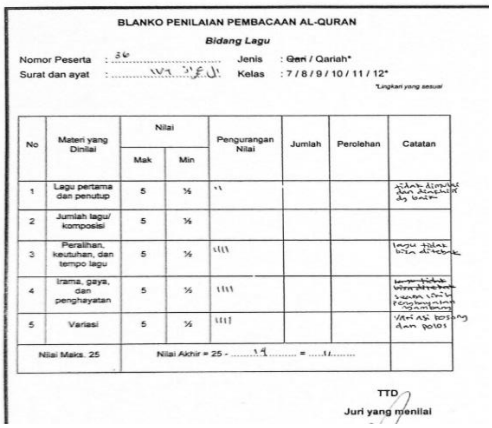
Catatan: suara lebih jelas suara kurang (terus) kurang jelas).

No	Materi yang Dinilai	Nilai Mak	Nilai Min	Pengurangan Nilai	Jumlah	Perolehan	Catatan
1	Lagu pertama dan penutup	5	½	1	1	4	
2	Jumlah lagu/ komposisi	5	½	1	3	2	Lag. kurang jelas
3	Peralihan, keutuhan, dan tempo lagu	5	½	1	3	2	
4	Irama, gaya, dan penghayatan	5	½	1	3	3	
5	Variasi	5	½	1	3	3	
Nilai Maks. 25		Nilai Akhir = 25 - 12 = 13					

TTD

Juri yang menilai

Rater AM


BLANKO PENILAIAN PEMBACAAN AL-QURAN

Bidang Lagu

Nomor Peserta : 36 Jenis : ~~Qari~~ / Qariah*

Surat dan ayat : Al-'Alaq 1-5 Kelas : 7 / 8 / 9 / 10 / 11 / 12*

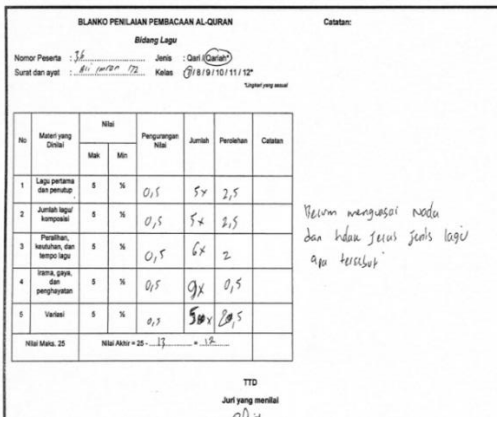
*Lingkari yang sesuai

No	Materi yang Dinilai	Nilai Mak	Nilai Min	Pengurangan Nilai	Jumlah	Perolehan	Catatan
1	Lagu pertama dan penutup	5	½	"			tidak dimulai dan berakhir di waktu
2	Jumlah lagu/ komposisi	5	½				
3	Peralihan, keutuhan, dan tempo lagu	5	½	||||			lagu tidak bisa di ketuk
4	Irama, gaya, dan penghayatan	5	½	||||			~~lagu tidak bisa di ketuk~~ lagu tidak berirama dan berlagu
5	Variasi	5	½	||||			Variasi tidak bisa dan tidak
Nilai Maks. 25		Nilai Akhir = 25 - 19 = 11					

TTD

Juri yang menilai

Rater AP


BLANKO PENILAIAN PEMBACAAN AL-QURAN

Bidang Lagu

Nomor Peserta : 36 Jenis : Qari (Qariah)*

Surat dan ayat : Al-'Alaq 1-2 Kelas : 7 / 8 / 9 / 10 / 11 / 12*

*Lingkari yang sesuai

Catatan:

No	Materi yang Dinilai	Nilai Mak	Nilai Min	Pengurangan Nilai	Jumlah	Perolehan	Catatan
1	Lagu pertama dan penutup	5	½	0,5	5x	2,5	
2	Jumlah lagu/ komposisi	5	½	0,5	5x	2,5	
3	Peralihan, keutuhan, dan tempo lagu	5	½	0,5	6x	2	
4	Irama, gaya, dan penghayatan	5	½	0,5	9x	0,5	
5	Variasi	5	½	0,5	5x	2,5	
Nilai Maks. 25		Nilai Akhir = 25 - 13 = 12					

Harus menguasai nada dan tahan jenis jenis lagu apa tersebut

TTD

Juri yang menilai

Figure 4. 19 *Qualitative report on the ratee number 36*

Evidence from the unexpected responses report

Further evidence related to the fairness aspect can be seen in the MFRM analysis results, specifically in the unexpected responses report. Tables 4.16, 4.17, and 4.18 are a synthesis of unexpected response information, grouping it by facet to demonstrate the MFRM's ability to identify unusual patterns in the assessment. The *Tajwed* dimension had 2.6% (21 out of 800) of unexpected responses, the *Fashahah* dimension had 3.8% (37 out of 960), the *Lagu* dimension had 4.2% (42 out of 1000), and the *Suara* dimension had 5.3% (53 out of 1000). These percentages are quite low compared to the total responses in each dimension. Therefore, in general, the rater can be considered to have given a careful assessment. Nevertheless, all raters were recorded as having given unexpected assessments to almost all ratees across all items in each dimension, except item F5, which was not included in the analysis. These findings show that MFRM analysis is very sensitive to the unusual responses that do not follow the assessment expectations. This ability indirectly

demonstrates the fairness or unfairness pattern in Quranic recitation assessment.

Table 4. 16 *Frequency of unexpected response distribution across items*

Items	Frequency	Response positive *(Observed>expected)*	Response negative *(Observed<expected)*
T1	1	0	1
T2	3	0	3
T3	7	0	7
T4	10	1	9
F1	9	0	9
F2	13	2	11
F3	12	0	12
F4	3	0	3
F5	-	-	-
L1	4	0	4
L2	10	1	9
L3	10	4	6
L4	12	3	9
L5	6	0	6
S1	6	1	5
S2	7	0	7
S3	17	0	17
S4	14	0	14
S5	9	8	1

Table 4. 17 *Frequency of unexpected response distribution across raters*

Rater	Frequency	Response positive *(Observed>expected)*	Response negative *(Observed<expected)*
SD	2	1	1
NA	8	0	8
AY	3	0	3
MA	8	0	8

SW	14	2	12
QN	13	0	13
RI	4	0	4
YN	6	0	6
NI	7	3	4
PB	17	1	16
AM	13	4	9
AP	5	0	5
UM	7	0	7
MF	24	4	20
JM	8	2	6
WM	14	3	11

Table 4. 18 *The frequency of the unexpected response for ratees across dimensions*

Freq	**Ratee in *Tajwed****	**Ratee in *Fashahah****	**Ratee in *Lagu****	**Ratee in *Suara****
1	1, 6, 8, 9, 10, 14, 16, 18, 19, 24, 29, 38, 42, 44, 45, 46, 47	2, 3, 8, 9, 11, 12, 15, 17, 18, 22, 24, 28, 29, 30, 32, 33, 34, 37, 38, 39, 43, 44, 45, 49	2, 5, 6, 9, 11, 12, 16, 18, 20, 22, 27, 31, 32, 34, 35, 45, 47, 49	2, 3, 6, 7, 14, 16, 17, 18, 25, 26, 28, 29, 31, 35, 36, 37, 39, 41, 43, 44, 46, 47, 48, 49,
2	5, 22	4, 7, 27	7 (+:1, -: 1), 10, 37, 48	4, 10, 13, 15, 27 (+:1, -: 1), 33, 42 (+:1, -: 1)
3	-	10, 50	39 (+:1, -: 2), 50 (+:1, -: 2)	1, 19 (+:1, -: 2), 20, 30, 40 (+:1, -: 2)
4	-	-	44 (+: 1, -: 3)	-
6	-	-	36 (+: 4, -:2)	-
Total	21	37	42	53
Positive Response	1	2	8	8

Negative response	20	35	34	45

*The number with the underline: The dominant ratee's responses were negative.

In general, unexpected response reports are consistent with the dominance of negative responses to almost all items. A negative response means that the observed score is lower than the score expected by the model (Linacre, 2023). In the context of unexpected responses, this scoring pattern is considered unusual according to the model. The *Lagu* and *Suara* dimensions are the dimensions that have the most negative responses in percentage. Referring to the MTQ assessment guideline, these two dimensions assess aesthetic aspects that require the assessor's subjective opinion, regardless of the existence of a *nagham* science that could serve as a guide (MTQ Guidebook, 2023). For instance, item L4 and item S3, as the items that have the most negative responses. Item L4 is rhythm, style, and appreciation, which the score depends on the rater's perception, which, according to this study's assumption, is based on the rater's knowledge and experience. If their capacity is different, the score may be different. Likewise, item S3, which assesses the smoothness and softness of the voice, also depends on the rater's subjectivity. Furthermore, item S5, which has the most positive unexpected responses, assesses breathing regulation, which can be heard and counted directly. However, there are still unexpected responses to this item. According to the MTQ assessment guideline book, there is no specific definition explaining each score or when each ratee gets the highest or the lowest score. Thus, even though the practice of breathing regulation can be heard clearly, the absence of clear descriptors opens the possibility of unexpected responses. It is not surprising that unexpected responses often occur in these two dimensions.

Regarding this study's assumption on the variation that occurs because the interpretation range is flexible, it can be explained by one of the assessments given by raters to the ratee number 36 in the *Lagu* dimension. As a ratee with the most unexpected responses (n: 6x), all raters in the *Lagu* dimension have given an unexpected response at least once to ratee number 36. Although previously qualitative reports have shown that all raters have the same understanding, it turns out that the responses given varied, some were lenient, and some were severe (Table 4.18). Not only that, highlighting the ratee's fit status reveals that the outfit MnSq value and the infit and outfit ZSTD are misfits, indicating that the responses are too unpredictable (Engelhard & Wind, 2018; Linacre, 2002). Failure to meet the ideal outfit value confirms that the unpredictable assessment pattern of ratee number 36 occurred due to the outlier factor, in this case, carried out by the rater.

According to The Standards (2014), *"When validity evidence includes statistical analyses of test results .... the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions (Standard 1.10, p. 36)"*. Highlighting the demographic background, the raters on the *Lagu* dimension are indeed quite varied. The demographic code for the NI rater is L1FSulut21A6PaTilTah, the PB rater: L2MGoron20TO6PaPeTilTahSyarFah, the AM rater: L3FJatim20US2PaPeTil, and the AP rater: L4MBanten20A3PaJuTilTah. From the information that sequentially explains the rater number, gender, province of origin, age, position in the *Pesantren*, length of stay in the *Pesantren*, experience in the MTQ competitions, and expertise in the MTQ competitions, only the expertise background crosses over, but, still, it is not exclusive because each rater has other expertise that provides new insights into the interpretation of the assessment in the *Lagu* dimension. Therefore, it is not surprising that the standardization of "Good or not good" on the items being assessed may be interpreted differently by each rater. However, this analysis is only an assumption that requires further examination.

Highlight the item fit condition; these four items have different statuses. Items L3 and L4 are detected to be suitable and have acceptable values, indicating that there are no issues with the items. However, in items L1 and L2, the value of infit MnSq (1.54 and 1.75, respectively) and Infit ZSTD (2.0 and 3.7, respectively) are detected as misfit (see table 4.5). This indicates that these two items are unproductive for measurement but not disturbing, and the data is too unpredictable (Engelhard & Wind, 2018; Linacre, 2002). Meanwhile, the findings of the disordered threshold and scale misfit in the *Lagu* dimension, especially in the category 2 and 3 scales, may also be the reason why the assessment in this dimension has many unexpected responses. Quoted from The Standards (2014), "*Characteristics of the test itself that are not related to the construct being measured, or the manner in which the test is used, may sometimes result in different meanings for scores earned by members of different identifiable subgroups*" (The Standards, 2014, p. 61). This statement reinforces the point that raters may respond inconsistently or unpredictably due to a lack of clarity in the construct, particularly in the item assessment descriptors or category scales. Therefore, improvements of the rubric must be made to ensure the assessment remains fair.

Table 4. 19 *Rater unexpected responses to the ratee number 36*

Ratee	Rater	Item	Observed	Expected
36	NI	L4	4	1.6
36	NI	L3	3	1.4
36	PB	L4	3	1.6
36	PB	L2	2	3.7
36	AM	L2	5	3.1
36	AP	L1	2	4.2

The next example of the unexpected response comes from the *Suara* dimension. The response from the raters to the item S5 for ratee number 3. Item S5 is an item that assesses breathing regulation. According to the MTQ book guidelines (2023), taking a breath in the middle of reading is a mistake because it can change the meaning of the verse being read. Unfortunately, one assessment pattern given by the rater MF is the opposite of what it should be. MF gave the following notes: "*Sering tanafus* (often *tanafus)*". Rater WM also confirmed with the notes, "*Suara kurang jelas, banyak tanafus, suara tidak jernih* (Vocals are not clear, many *tanafus*, voice is not clear)". Meanwhile, Raters UM and JM, although they did not provide qualitative notes, reduced the score of item S5 for ratee number 3. This differs from rater MF, who already understood that the ratee made a mistake but still gave the maximum score. This suggests that MF was probably careless and entered a score that did not match expectations.

To provide a comparative analysis of the *Lagu* and *Suara* dimensions, which emphasize aesthetic perception, examples of assessments on the *Tajwed* and *Fashahah* dimensions will also be explored. Like the aesthetic dimension, which is assumed to provide variation due to the data fit, bias/interaction report, and the rater's background, assessments on dimensions that consider theory in reciting the Qur'an also provide variation due to similar reasons. For example, in ratee number 8, rater NA was recorded as giving a negative response, indicating that the observed value differed from the expected value by the model, which was 1 out of 4. According to the qualitative notes, rater NA provided more comments than the other raters. This study presents the following details about the location of the notes in various forms, specifically in the *maqra* of the Quran, surah At-Taubah, verses 43 to 47.

عَفَا اللّٰهُ عَنْكَ لِمَ اَذِنْتَ لَهُمْ حَتّٰى يَتَبَيَّنَ لَكَ الَّذِيْنَ صَدَقُوْا وَتَعْلَمَ الْكٰذِبِيْنَ ٤٣

لَا يَسْتَاْذِنُكَ الَّذِيْنَ يُؤْمِنُوْنَ بِاللّٰهِ وَالْيَوْمِ الْاٰخِرِ اَنْ يُّجَاهِدُوْا بِاَمْوَالِهِمْ وَاَنْفُسِهِمْ ۗوَاللّٰهُ عَلِيْمٌۢ بِالْمُتَّقِيْنَ ٤٤

اِنَّمَا يَسْتَاْذِنُكَ الَّذِيْنَ لَا يُؤْمِنُوْنَ بِاللّٰهِ وَالْيَوْمِ الْاٰخِرِ وَارْتَابَتْ قُلُوْبُهُمْ فَهُمْ فِيْ رَيْبِهِمْ يَتَرَدَّدُوْنَ ٤٥

۞ وَلَوْ اَرَادُوا الْخُرُوْجَ لَاَعَدُّوْا لَهٗ عُدَّةً وَّلٰكِنْ كَرِهَ اللّٰهُ انْۢبِعَاثَهُمْ فَثَبَّطَهُمْ وَقِيْلَ اقْعُدُوْا مَعَ الْقٰعِدِيْنَ ٤٦

لَوْ خَرَجُوْا فِيْكُمْ مَّا زَادُوْكُمْ اِلَّا خَبَالًا وَّلَاَوْضَعُوْا خِلٰلَكُمْ يَبْغُوْنَكُمُ الْفِتْنَةَ ۚوَفِيْكُمْ سَمّٰعُوْنَ لَهُمْ ۗوَاللّٰهُ عَلِيْمٌۢ بِالظّٰلِمِيْنَ ٤٧

In these verses, one colour is particularly dominant, namely green. Green is the note owned by the NA rater, while yellow is the MA rater, blue is for the AY rater, and purple is for the SD rater. This note is specifically given for item T3, namely *Ahkam al-hurf,* which assesses the accuracy of the *Tajwed* recitation law, such as *idzhar, ikhfa, idgham*, or *iqlab.* According to the model, the NA rater made an unexpected response, which should have resulted in a score of 4, but the NA rater only assigned a score of 1. This result is reasonable because the scores given between the NA rater and the other three raters are quite different, namely: SD: 2, AY: 3, MA: 3. This number does not represent the number of errors recorded by the rater, but rather the score converted based on the likert rating scale rules in this study (Table 3.7). The expected score difference can also be explained in the NA rater statistical report. Compared to the other raters, the NA's distance from the model's expected agreement is the widest at 6.7%. The infit ZSTD value of this rater is also outside the ideal range, +2.4, which according to Linacre (2002) is considered to have unproductive measurement, even though not disturb the measurement.

Highlight the demographic background of the rater, the code for each rater is as follows: SD rater: T1FSulut29T11PaJuPeTil, NA rater: T2Bali21TOPaPeTilTahSyarFahMurQir, AY rater: T3MGoron23TO7PaPeTilTahQir, and MA rater: T4SulSel19TO1PaPeTilMurQir. The striking difference is indeed quite visible in the rater's area of expertise in the *Tajwed* dimension. Rater NA is recorded as having six expertise in the MTQ competition field, namely *Tilawah, Tahfidz, Syarhil, Fahmil, Murottal,* and *Qiroat*. This number is indeed a lot when compared to the others. This study assumes that the amount of expertise possessed may have influenced the rater's strictness

or accuracy in assessing the rater's performance. Yet, this is just an assumption that needs further examination.

As for unexpected responses in the *Fashahah* dimension, they are quite diverse. It is difficult for this study to explain a specific example because it is complicated to find qualitative data that support each other. Along the way, this study found many rater notes in the *Fashahah* dimension that mentioned comments related to other dimensions. Such as comments related to the law of *Tajwed*, the law of *Mad*, and the practice of *tanaffus,* which should be noted by raters in the *Tajwed* and *Suara* dimensions (Figure 4.20). While some notes given by raters in the *Fashahah* dimension are relevant, the presence of qualitative data that should be noted in other dimensions suggests that raters may not understand the aspects that must be assessed in their dimensions. Look at the data fit on the rater and items; there are no significant problems except for item F4 and rater YN, which are considered too consistent or overfit according to the model because the outfit MnSq value they have is below standard. Highlighting the bias reports, all responses in this dimension are also recorded as not giving any bias at all. This study suspects that this condition occurred due to the descriptor of the item that was unclear and or due to the rater's abilities in differentiating the items questioned.

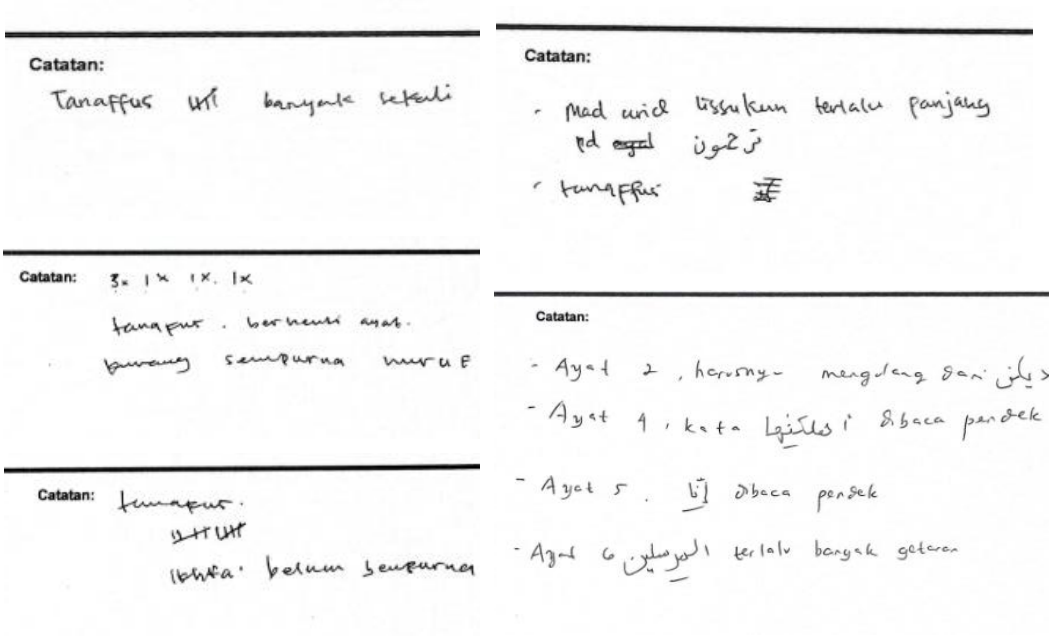


Figure 4. 20 *Qualitative report of raters in the Fashahah dimension*

Based on the report, it can be understood that the unexpected response conditions

occur quite rarely compared to the total response. The rater's careful assessment proves to be good, even though the model captures an unexpected pattern in which most raters tend to give positive responses, but under certain conditions show a negative or severe pattern. This condition certainly needs to be evaluated. Evidence of unfairness in the assessment has been found in several conditions, they are occurred due to the data misfit either cause by rater or ratee performance, the issues on the item descriptor or category scale errors (disordered threshold), the bias in the assessment that occur because the raters are too lenient or too severe, and the unexpected responses by raters to ratees on certain items. Finally, the findings above show that the assessment was conducted fairly, although unfairness in several conditions can still be identified to enhance the quality of the rubric.

4.3. Discussions

In line with the objectives of this study, the analysis in sections 4.1 and 4.2 has shown that the Quranic recitation assessment rubric has decent quality in terms of psychometric aspects: validity, reliability, and fairness. Although the rubric generally performed well, several findings raised critical notes, particularly on several indicators, scale structures, and variations in assessments between raters. To explain the meaning and implications of these findings, this section will elaborate further by referring to previous theories and studies.

Multi-rater-based assessment is a complex, yet very useful and meaningful tool for producing fair, robust, and comprehensive measurements. Many measurement practices that require the involvement of many experts have decided to utilize this method to make their judgments more valid. Several studies have adopted a multi-rater method to evaluate instruments in various fields. For instance, the study by Scherbakova et al. (2025) tested the Scale of Aesthetics and Creativity in Chess (SACC) instrument for a chess competition. On the other hand, Afifi et al. (2023) used this method to test the Quranic Verbal Communication Index (QVCI). Both studies employed MFRM analysis, which proved effective not only in validating the instruments but also in uncovering important psychometric properties in a multi-rater context.

MFRM has also become a widely accepted tool, particularly in performance-based assessment. For example, research by Arifiyanti et al. (2023), which examined bias in research seminar presentation evaluations, research by Maryati et al. (2019), which assessed teachers' teaching performance, and Aryadoust (2015), which assessed the oral presentation performance of first-year university students. Not only that, in the artistic or creative assessment context, where judgment is often highly subjective (Scherbakova et al.,

2025), MFRM can still be a useful tool, as long as the practice involves multi-rater feedback. In a competition held by the Creative Wood Workers Group (GSKK), Nor et al. (2024) provided a fairer assessment analysis. Springer et al. (2018) conducted a bias investigation in the evaluation of band concerts. Kudiya et al. (2018) conducted an aesthetic assessment of Batik artwork by Batik artisans. Wind et al. (2016) researched music performance assessment that provides valuable information for improving the accuracy and quality of assessments. As far as the researcher's research goes, many arts competition-based practices have also been studied, such as Alvarez-Diaz et al. (2021), who tested musical contest performances, Gynild (2016), who assessed vocal performances, Mitchell (2014), who evaluated singing voices, and Bahruddin and Khumaidi (2014), who assessed Quranic recitation competitions. However, the use of MFRM as an analysis tool for those kinds of competitions is still rarely found. Including one of the religious-based art competitions raised by Bahruddin and Khumaidi's research (2014), namely the practice of Quranic recitation at the Musabaqah Tilawatil Quran (MTQ) competition.

The MTQ competition is a performance-based competition involving multi-raters. It has experienced several allegations of fraud in the assessment and still uses traditional analysis tools in its process. Bahruddin and Khumaidi (2014) developed a new system of assessment model by critically evaluating the assessment system, which they claimed did not align with the principles of assessment and measurement. Through the model development procedure of Borg and Gall (1983), they conducted pre-development by highlighting the need to use the same *maqra'* between participants, the need to use a proportional assessment system, as the score reduction system has the potential to produce unfair assessments due to differences in *maqra'* between participants. They also identified the need for a proportional distribution of items' tasks on the *Tajwed* and *Fashahah* dimensions, as too many components have the potential to produce inconsistencies in assessments. Additionally, they identified the need to revise the rules for the limit of score differences, as these rules can cause the assessments to be impure and not objective. The results of those analyses are critical and reasonable, but no empirical evidence has been displayed explaining the impact of these practices on unfair assessment before the system is revised. Testing the rubric before it is revised is essential to understand its condition before revision and development (The Standards, 2014). Therefore, this study is presented to provide empirical evidence, starting with testing the Quranic recitation assessment rubric by utilizing a modern analysis tool, MFRM. A fundamental analysis of measurement is carried out by addressing the validity, reliability, and fairness of the rubric.

Validity of the rubric

In general, the construct validity aspect demonstrates that the indicators in the rubric accurately describe the aspects of the Quranic recitation assessment to be measured. This can be seen from the suitability of the data with the model and the indices of the indicators, most of which show a good fit. However, there is one item that is considered invalid and needs further attention. Item F5, which assesses punctuality in recitation (*Tamam al-waqt),* is in the *Fashahah* dimension. Compared to other items in this dimension, punctuality deviates slightly from the "*Fashahah*" aspect, which assesses the clarity of reciting *hurf, kalimah*, and *ayah* (MTQ Guidelines, 2023). If the consideration is whether a reader exceeds the time limit by adding a verse after time has ended, this aspect overlaps with the component of item F4, which assesses *Muo'ah al-ayah* by determining if any verses have been added or missed. These results indicate that item F5, in the case of this study, is considered less effective in measuring the *fashahah* aspect. The researcher assumes that the punctuality assessment aspect is more appropriately placed in the *Lagu* dimension; the reasons have been previously explained in section 4.2.1.

Examining the criticism given by Bahruddin and Khumaidi (2014), it is evident that one of the criticisms was directed towards the numerous components evaluated within the *Tajwed* and *Fashahah* dimensions. In line with this study, the *Fashahah* dimension was recorded as having an additional task: assessing punctuality in addition to the theoretical aspects of participants' fluency in reciting the Quran. According to Bahruddin and Khumaidi (2014), this condition has the potential to lead to inconsistencies in rater assessments. The similarity between F4 and F5 does not add new information, as evidenced by the absence of variation in item F5, which, according to this research, has overlap assessment descriptors. This condition further complicates the assessment process, as indicated by the low reliability and separation values in the *Fashahah* dimension. This is also indicated by the existence of construct-irrelevant variance, where raters cannot consistently map the participants' performance to the intended assessment item category (see Table 4.7, which explains the disordered threshold). One of the impacts can be seen in the inconsistency of the rater, which sometimes yields too predictable or too unpredictable scoring (Kassim, 2011). This is suspected to be due to the rater's confusion, as proved by qualitative notes that comment on aspects of other dimensions (see explanation in section 4.2.3).

Another critical note was found in the response process, which revealed many

response imbalances. Previously, researchers had to emphasize that the likert rating scale used in this study is a category scale created by the researcher for the need of MFRM analysis. The scale descriptors are spaced similarly so that the responses to the scale can be distributed well. However, the results of the analysis revealed many imbalances, particularly in the distribution of the low and high scale categories. According to the Rasch Andrich threshold theory, the failure to achieve a balanced condition is possible because the category scale overlaps or is too wide (Tennant, 2004; Linacre, 2023). For this reason, revision of the scale descriptor or restructuring the scale categories needs to be considered as the results of the statistical indicators explained in Tables 4.6 and 4.7. Regardless, the large number of category scores distributed in the upper group indicates that the ratees tend to perform well, and this is quite reasonable because 76% of the participants in this study had participated in the MTQ competition (see Table 3.1).

Those conditions confirm that the empirical approach using MFRM can reveal anomalies in constructs that a normative approach cannot detect alone. The critical statements given by Bahruddin and Khumaidi (2014) on the number of tasks in the *Fashahah* dimension can finally be tested empirically, though their opinion is not fully confirmed due to limitations in different research contexts. These results reveal that, in terms of validity, the Quranic recitation assessment rubric has decent construct validity, although some critical notes on several items and scale categories need to be considered to improve its validity.

Reliability of the rubric

In terms of reliability, the analysis results show that the rubric performs quite well in producing consistent scores between raters and between ratees. The evidence is based on the indicator values that demonstrate the consistency and precision of the measurements obtained in performance-based, assessment-based practices. However, the reliability and separation values of ratees in the *Fashahah* dimension tend to be low. This condition indicates a tendency for homogeneity in the ratee scoring pattern (Engelhard & Wind, 2018; Fisher, 2007). Low separation indicates that the ratees provide too consistent variation in their performance (Bond et al., 2021). Therefore, even though the reliability indices for other facets are in a good range, evaluations are still needed to improve their sensitivity and discrimination, particularly for the assessment of ratees' performances.

The low values of the reliability and separation of ratees in the *Fashahah* dimension confirm the findings in other aspects. First, the invalidity of item F5, according

to the dummy data report, was recorded as having the same scoring pattern across all ratees by all raters. The absence of variation in the scores given indicates that the model is unable to detect new information that can be reported by the software (Wind, 2018). The report on the response pattern of the scale category also shows a tendency to give high scores that are too unequal from those of other items in the *Fashahah* dimension. This further confirms that the model is having difficulty capturing the data variability of ratees.

That issue cannot be evaluated based solely on the ratee's performance. The overlapping scores suggest that the scale category descriptors are ineffective at capturing the scoring pattern (Tennant, 2004; Wind, 2018; Andrich, 2019). Qualitative rater notes on the *Fashahah* dimension reveal reporting anomalies, which suggests that the raters may contribute to the lack of variation in ratee scoring patterns. Anomalies suspected to be due to rater confusion regarding the items also indicate the need to revise the item descriptors so that the items can be understood by the raters. In addition, the absence of bias in the *Fashahah* dimension may be caused by insufficient data variation, meaning the model does not capture any deviant assessments. The tendency for high scores in the *Fashahah* dimension may indicate that the majority of ratees are indeed quite capable in the *Fashahah* aspect (Bond & Fox, 2015; Sumintono & Widhiarso, 2015), but the anomaly in the qualitative notes indicates that this assumption is not strong enough.

The finding of inconsistent ratee performance patterns with insufficiently varied data indicates the need for evaluation in various aspects. However, the non-implementation of one of the stages in the MTQ competition in this study, namely, the deliberation of assessment between raters to achieve a consistent score, ensures that this study's assessment was objective. This stage is not used because many researchers have stated that the assessment must be independent to maintain its objectivity (Rasch, 1960; Wright & Stone, 1979; Linacre, 1994; Engelhard, 2013; Sumintono & Widhiarso, 2015; Bond & Fox, 2015; Wind, 2018). Bahruddin and Khumaidi (2014), who studied in this rubric, also criticize the deliberation stage between assessors. The reason is that the result given is not from the original value but from the deliberation and negotiation. Mitchell (2014), who assessed singing voices, and Wind et al. (2016), who assessed music performance, stated that subjectivity, which can lead to the freedom of raters' judgment, is unavoidable. Wang et al. (2020) also stated that it is unrealistic if the assessment requires all raters to produce identical assessments. Those confirmed by the research of Gynnild (2015), who assessed vocal performances, that different opinions were normal to take a place. Therefore, instead of conducting the deliberation, sharpening item descriptions, rater training, or tightening

the selection of raters are solutions that this study offers. The idea came from Alvariez-Diaz et al. (2021), who found disparities in the assessment of musical contest performance, and sharpening rubric descriptors was the solution. Ogari (2020) in his dissertation also suggested standardizing a clear assessment rubric as the solution offered.

However, even though an assessment is a practice that must prioritize objectivity and independence, evaluating the systems is the fair solution to maintain the consistency of the assessment. By utilizing MFRM analysis, the data scoring can be calibrated and estimated to reveal the anomaly in assessment (Wang et al., 2022). The findings on the low indicator of the ratee group in the *Fashahah* dimension are an example of how a fair solution can be found. Here, the rubric needs improvement, particularly regarding its sensitivity in discriminating the pattern of how ratees are assessed or raters assess the Quranic recitation performance. The goal is to make the practices that occur more precise and consistent in various conditions.

Fairness of the rubric

The fairness in performance-based assessment is a crucial component that is often overlooked in the development of assessment rubrics. The results of the analysis show that the Quranic recitation assessment rubric generally shows a good level of fairness. Most data reports fit the model; the potential bias only occurs in 1.4% of interactions, and only 0.02% of interactions are statistically significant. This indicates that the rubric can facilitate an assessment system that is relatively free from systematic bias. However, the emergence of several unexpected responses, albeit in small numbers, is a cause for concern.

Unexpected response is an assessment pattern that deviates from the expected pattern (Linacre, 2023). Previous researchers have found that this condition stems from various factors, such as inconsistency of the rater, problems with the rubric, either due to unclear descriptors or ineffective category scales, and psychological or technical conditions of the rater or ratee during the assessment process (Arifiyanti et al., 2023; Wang et al., 2020; Wind, 2018; Aryadoust, 2016; Kassim, 2011; Moskal & Leydens, 2002). Internally, the misfit or overfit of some items with the model and the disordered threshold on the middle scale in almost all dimensions indicate that rater inconsistency may occur due to their confusion with unclear item descriptors and inappropriate category designs. To overcome this, Moskal and Leydens (2002) suggested revising the item descriptor by ensuring clarity of the descriptor, clearly differentiating, being consistent in language and avoiding quantifying words such as ‘less’, ‘some’, and Wang et al. (2020) added that providing

training for raters can also be done to reduce unexpected responses. Meanwhile, revisions to the category design that cause disordered thresholds must be done by combining or adding category scales to be more sensitive in measuring participant performance (Wind, 2018).

Externally, this study found that the raters' scientific background is assumed to influence the consistency and strictness of their assessment. In some conditions, raters who have more experience tend to provide more detailed notes and more severe assessments (see the pattern of NA rater assessments to ratee number 8). The perception of scoring is also different in the example of rater assessments in the *Lagu* dimension on ratee number 36, which is assumed to occur due to differences in the rater's scientific background. This finding was revealed in the research of Arifiyanti et al. (2023) and Aryadoust (2016), who both assessed a research seminar presentation, that scientific background influences and has the potential to bias the assessment. This finding makes it clear that regular training for raters is important so that their perceptions during the assessment are held to the same standard.

The findings on the influence of internal and external factors explain that in the performance-based assessment process, the element of subjectivity is one thing that cannot be avoided. However, it does not mean that it can be left alone. To create a fair assessment, the solution provided must also support the principles of valid, reliable, and objective measurement. For that, the best solution in overcoming the fairness aspect is to create the best possible rubric by considering the ecological aspects that may affect the quality of the assessment. In the context of Quranic recitation assessment, the rubric must be compiled systematically, fairly, and within its portion and capacity. Ecological aspects outside the facets must also be recognized, but at the same time must be overcome by providing an opportunity to learn for the rater as the assessor, the ratee as the performer, and even the item as a component of the assessment to be continuously developed. This point is needed to emphasize because with the invalid, unreliable, or unfair assessment in this competition, the result of the assessment, which is used to determine the winner is possibly not represent the best performance in the Quranic recitation. Therefore, as stated in The Standards (2014), *"Opportunity to learn the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test has several implications for the fair and valid interpretation of test scores for their intended uses (p. 66)."*

Finally, quoting from the research conclusions given by Kassim (2011), "*We may not be able to eliminate rater errors, but we can minimize them in some ways. The use of a robust measurement model is one. Rater training is another. But the most important is the human factor itself. No amount of rater training can change a poor attitude, and no measurement model can correct for inconsistent and poor ratings. If we are not willing to make the effort to ensure that our judgment of our students' performance is reliable and valid, nothing else can be done. As assessment and its procedures are at the heart of student learning, matters related to valid and fair testing need to be taken seriously. It is hoped that with greater awareness of how we judge, we can be better raters and better teachers*." And that is the meaning of "Fairness" as mentioned in the Quran, surah Al-Maidah verse 8, "*O believers! Stand firm for Allah and bear true testimony. Do not let the hatred of a people lead you to injustice. Be just! That is closer to righteousness. And be mindful of Allah. Surely Allah is All-Aware of what you do*."

CHAPTER V
CONCLUSION

This chapter is the last part of this study, which contains the conclusion, limitations, and recommendations of the study. The conclusion is the section that summarizes all the findings on each research question. Limitations are the section that explains the shortcomings of this study, which then becomes the basis for recommendations for academicians, practitioners, and readers in the future. All these results are expected to provide scientific contributions and significance, especially in improving the quality of the Quranic recitation assessment rubric instrument and offering the benefits of Many Facet Rasch Measurement as an advanced multi-rater analysis tool.

5.1. Conclusion

This study evaluates the Quranic recitation assessment rubric used in the MTQ competition. The focus is on three psychometric aspects: validity, reliability, and fairness. The MFRM analysis approach was chosen because it can analyse performance-based assessment data involving many raters. The results summary for each research question is provided as follows:

First, in terms of validity. The Quranic recitation assessment rubric is declared valid because most indicators show their suitability with the model. Evidence provided on the internal structure and response process shows acceptable results. However, there is one item that needs to be revised or deleted in the *Fashahah* dimension because it is not effective in measuring the aspects that should be measured. Other findings on disordered thresholds in several scale categories indicate the need for revision of the scale descriptor, or scale restructuring that needs to be collapsed or added to be more sensitive in discriminating against the level of participant ability.

Second, in terms of reliability. The assessment provided through the Quranic recitation assessment rubric is considered reliable. This is supported by the measurement index report, which shows good consistency and accuracy across raters, items, and ratees. However, there are notes about the performance patterns of ratees in the Fashahah dimension because they demonstrate low consistency and discrimination values in assessment. These results suggest that the ratees' performances do not provide new information for evaluating their quality. Nonetheless, this condition remains within acceptable ranges and can still be improved.

Third, in terms of fairness. The assessment conducted in this study was declared statistically fair. Most raters and ratees showed fit to the model. Only one interaction was found to be significantly biased out of 3.760 interactions. Some unexpected responses did appear but were still within reasonable limits. Several factors contributed to this, including the ineffectiveness of item descriptors and category rating scales, internal rater factors such as fatigue and confusion, which led to carelessness in assessments, and external rater factors like scientific backgrounds that influenced their perceptions and scoring standards. These findings indicate that although the assessment was fair from a technical perspective, true justice also depends heavily on the personal integrity and ethical awareness of the raters. Therefore, the MFRM approach is not only able to reveal the technical quality of the rubric but also opens space for awareness that ecological aspects also have an important influence on the performance-based assessment process, in this case, Quranic recitation performance.

5.2. Limitations

This study has several limitations that need to be considered. First, the research context in this study is limited, so the findings provided cannot be generalized widely. Second, although the MFRM approach offers an in-depth quantitative analysis, the assumption that rater demographic factors influence the assessment results cannot be proven empirically due to the limited method and analysis tools employed in this study. Third, this study does not consider the other facets, which are the *maqra'* or assignment of the ratees in their Quranic recitation performance. Not only that, but the topic also explored in this study is quite specific, which may not be widely known to some audiences.

5.3. Recommendations

Based on the findings and considering the limitations of this study, several recommendations are given to related parties to develop a Quranic recitation assessment rubric, as well as to the academic community regarding performance-based assessment and its rubric assessment development.

a. MTQ organizers and activists

MTQ organizers, in this case LPTQ, are recommended to periodically evaluate and develop assessment rubrics to remain relevant with the dynamics of participant quality that continues to improve. Openness to the use of modern analysis tools such as Many Facet Rasch Measurement (MFRM) also needs to be considered so that the assessment process becomes more objective and data-based. In addition, intensive training for judges also

needs to be held to improve consistency, reduce bias, and strengthen the fairness of raters in giving assessment scores.

b. Recommendations for future scholars and practitioners

The next recommendation is given to scholars and practitioners of educational evaluation, but not limited to, to expand the scope of the study. Future researchers can continue this research, such as expanding the scope of the study sample so that the results can be more widely generalized. Correlational studies and the development of empirical evidence-based rubrics can also be further steps to deepen the study of the Quranic recitation assessment rubrics' quality in the MTQ competition. On the other hand, the dissemination of Rasch-based measurement literacy into wider fields can also be done to continue to develop and contextualize Rasch measurement.

c. Educational institutions offering Quranic Recitation courses

For educational institutions such as Islamic Boarding Schools or *Pesantren*, using MFRM in the student evaluation process is recommended. Through this method, it allows teachers and students to obtain more detailed, measurable, and fair evaluative feedback, thus the coaching process can be carried out more precisely. In addition to improving the learning quality, MFRM can also help their students with better preparation for the MTQ competition.

d. Recommendations for the Rasch measurement community

Further recommendations are also given to the scientific community that focuses on Rasch measurement and its branches, including MFRM. This study has proven that MFRM is not only relevant for general education but can also be applied effectively in assessing religious and aesthetic-based performance, such as Quranic recitation. This shows new opportunities for the diversification of Rasch application in a wider social and cultural context.

REFERENCES

AERA, APA, and NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Afifi, S., Kurniawan, I. N., and Sumintono, B. (2023). Pengembangan instrumen penelitian the Qur'anic Verbal Communication Index (QVCI) menggunakan pemodelan rasch. *Jurnal Ilmu Komunikasi, 21*(1), 94–112. DOI: https://doi.org/10.20885/komunikasi.vol21.iss1.art6

Al-Azhar University. (n.d.). About Al-Azhar University. Retrieved from *https://www.azhar.edu.eg*

Ali, M. (2016). Penerapan teknologi e-MTQ pada MTQ nasional. *Jurnal Multikultural & Multireligious, 15*(3), 143-160.

Ali, M. I. (2024). A guide for positivist research paradigm: From philosophy to methodology. *Ideology Journal, 9*(2), 187-196. DOI: https://doi.org/10.24191/idealogy.v9i2.596.

Al-Jazari, M. (2017). *Al-Muqaddimah al-jazariyyah: Introduction to the science of tajwed.* Riyadh: Darussalam.

Al-Mahalliy, J., and Al-Suyuti, J. (1990). *Terjemah tafsir jalalain berikut azbabun nuzul,* trans. Bakar, A.B. Bandung: Sinar Baru Officer.

Al-Nassir, A. A. (1985). *Sibawayh the phonologist: A critical study of the phonetic and phonological theory of Sibawayh as presented in his treatise al-kitab.* Dissertation. University of York.

Al-Qattan, M. K. (1995). *Mabahis fi ulumil Quran*. Kairo: Maktabah Wahbah.

Alvarez-Diaz, M., Muñiz-Bascón, L. M., Soria-Alemany, A., Veintimilla-Bonet, A., and Fernández-Alonso, R. (2021). On the design and validation of a rubric for the evaluation of performance in a musical contest. *International Journal of Music Education, 39*(1), 66–79. DOI: https://doi.org/10.1177/0255761420936443

Amrullah, A.M.A.K. (2001). *Tafsir al-azhar*. Singapura: Pustaka Nasional Pte. Ltd.

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27–31. DOI: https://doi.org/10.3200/CTCH.53.1.27-31.

Andrich, D., and Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social, and health sciences*. Singapore: Springer Nature.

Andrich, D. (1988). *Rasch models for measurement.* Sage Publications, Inc.

Anfara, V. A., and Mertz, N. T. (2006). *Theoretical frameworks in qualitative research*. SAGE Publications.

Arifiyanti, F., Soeharto, S., Amukune, S., Van Nguyen, S., Aburezeq, K., Hidayatullah, A., and Sarimanah, E. (2023). Investigating rater-student interaction, gender bias, and major bias in the assessment of research seminar presentations. *Heliyon*, *9*(6), e16548. https://doi.org/10.1016/j.heliyon.2023.e16548

Aryadoust, V. (2015). Self- and peer assessments of oral presentations by first-year university students. *Educational Assessment, 20*(3), 199-225. DOI: 10.1080/10627197.2015.1061989

Aslanoğlu, A.E., and Şata, M. (2021). Examining the differential rater functioning in the process of assessing writing skills of middle school 7th grade students. *Participatory Educational Research, 8*(4), 239-252. https://doi.org/10.17275/per.21.88.8.4

Azwar, A.F. (2018). Gagasan rekonstruksi tradisi Musabaqah Tilawatil Quran (MTQ) dalam perspektif rahmatan lil 'alamin. *Jurnal Ilmu Agama: Mengkaji Doktrin, Pemikiran, dan Fenomena Agama, 19*(1). DOI: https://doi.org/10.19109/jia.v19i1.2379

Badan Pusat Statistik. (n.d.). Bridging kode. Retrieved January 6, 2025, from *https://sig.bps.go.id/bridging-kode/index*

Baharuddin, D., Jamaa, L., and Syarif, R. A. (2022). The Qur'an in a Christian majority: A case study of tolerance in the 29th MTQ in Saumlaki, the Moluccas, Indonesia. *Wawasan: Jurnal Ilmiah Agama Dan Sosial Budaya 7*(2), 121-130. DOI: https://doi.org/10.15575/jw.v7i2.22631.

Bahrudin and Kumaidi. (2014). Model asesmen musabaqah tilawah al-quran (MTQ) cabang tilawah, *Jurnal Penelitian dan Evaluasi Pendidikan*, *18*(2), 153-167. DOI: https://doi.org/10.21831/pep.v18i2.2858.

Bond, T. G., and Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd Ed.)*. Routledge.

Bond, T., Yan, Z., and Heene, M. (2021). *Applying the rasch model: Fundamental measurement in the human sciences* (4th Ed.). Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J., Staver, J. R., and Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Booth, A., Sutton, A., and Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (2nd Ed.). SAGE Publications.

Brooker, S., and Antonini, A. (2025). On the aesthetics of hypertext: a case study on names and a general framework. *New Review of Hypermedia and Multimedia, 31*(1–2), 7–29. https://doi.org/10.1080/13614568.2025.2463889

Brookhart, S., & Guskey, T.R. M. (Eds.). (2019). *What we know about grading: What works, what doesn't, and what's next*. ASCD.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, VA: ASCD.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M.T., Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848. DOI: https://doi.org/10.3102/0034654316672069.

Brunswick, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Bryman, A. (2016). *Social research methods* (5th Ed.). Oxford University Press.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly, 12*(3), 262–282. DOI: https://doi.org/10.1080/15434303.2015.1053134

Carlson, M. D. A., and Morrison, R. S. (2009). Study design, precision, and validity in observational studies. *Journal of Palliative Medicine, 12*(1), 77–82. https://doi.org/10.1089/jpm.2008.9690

Carmona-Halty, M., Salanova, M., Llorens, S., and Schaufeli, W. B. (2021). Linking positive emotions and academic performance: The mediated role of academic psychological capital and academic engagement. *Current Psychology, 40*, 2938–2947. https://doi.org/10.1007/s12144-019-00227-8.

Connelly, B.S., Warren, R.A., Kim, H., and Di Domenico, S.I. (2016). Development and validation of research scales for the leadership multi-rater assessment of personality (LMAP). *International Journal of Selection and Assessment, 24*(4), 362–367. https://doi.org/10.1111/ijsa.12154

Cooksey, R.W. (1996). *Judgment analysis: Theory, methods and applications*. Bingley, UK: Emerald. https://archive.org/details/judgmentanalysis0000cook

Creswell, J. W., and Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th Ed.). SAGE Publications.

Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, INC.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. DOI: https://doi.org/10.1007/BF02310555.

Dawson, P., and Dobson, S. (2010). The influence of social pressure and nationality on individual decisions: Evidence from the behaviour of referees. *Journal of Economic Psychology, 31*(2), 181-191. DOI: https://doi.org/10.1016/j.joep.2009.06.001.

Dubai International Holy Quran Award. (2020). *Annual report of DIHQA*. Dubai: Dubai International Holy Quran Award Organization.

Eckes, T. (2011; 2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments.* New York: Peter Lang.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112. DOI: https://doi.org/10.1111/j.1745-3984.1994.tb00436.x.

Engelhard, G. (2002). *Monitoring raters in performance assessments*. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781410605115-13/monitoring-raters-performance-assessments-george-engelhard

Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., Jr., and Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Engelhard, G., Jr., and Wang, J. (2024). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences* (2nd Ed.). Routledge. DOI: https://doi.org/10.4324/9781003458746

Engelhard, G., Jr., and Wang, J. (2021). *Rasch model for solving measurement problems: Invariant measurement in the social sciences.* Sage Publications.

Engels, M. C., Spilt, J. L., Denies, K., and Verschueren, K. (2021). The role of affective teacher-student relationships in adolescents' school engagement and academic achievement. *Journal of Educational Psychology, 113*(5), 930–944. https://doi.org/10.1037/edu0000493.

Fraenkel, J. R., Wallen, N. E., and Hyun, H. H. (2012). *How to design and evaluate research in education.* McGraw-Hill.

Fisher, W. P. Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1). https://www.rasch.org/rmt/rmt211m.htm

Gani, N. A. A. (1989). *Peristiwa dan sejarah kelahiran MTQ pertama*. Yayasan MTQ.

Gibbons, F. X. (2015). *Psychology of evaluation: Affective processes in cognition and emotion.* Springer. DOI: https://doi.org/10.4324/9781410606853

Ginkel, S.V., Laurentzen, R., Mulder, M., Mononen, A., Kyttä, J., and Kortelainen, M. J. (2017). Assessing oral presentation performance: Designing a rubric and testing its validity with an expert group. *Journal of Applied Research in Higher Education 9*(3), 474-486. DOI: https://doi.org/10.1108/JARHE-02-2016-0012.

Gynnild, V. (2016). Assessing vocal performances using analytical assessment: A case study. *Music Education Research, 18*(2), 224–238. DOI: https://doi.org/10.1080/14613808.2015.1025732

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Han, C. (2021). Detecting and measuring rater effects in interpreting assessment: A methodological comparison of classical test theory, generalizability theory, and many-facet Rasch measurement. In *Testing and assessment of interpreting*, pp. 85–113. Springer. DOI: https://doi/10.1007/978-981-15-8554-8_5

Hart, C. (2018). *Doing a literature review: Releasing the social science research imagination* (2nd Ed.). SAGE Publications.

Hasan, A. R. (2019). Pendidikan karakter bersaing dalam Musabaqah Tilawatil Qur'an. Ilmu Al-Qur'an. *Ilmu Al-qur'an, 2*(2), 202-216. DOI: https://doi.org/10.37542/iq.v2i02.33.

Armiadi, F., Khairul Nuzuli, A., & Oktavia, W. (2023). Potensi keagamaan pada anak dan remaja melalui program MTQ, mahasiswa kuliah kerja nyata di Nagari Batang Arah Tapan. *AMMA: Jurnal Pengabdian Masyarakat*, *2*(7), 882–890. Retrieved from https://journal.mediapublikasi.id/index.php/amma/article/view/3509

Hogarth, R. (1987). *Judgment and choice: The psychology of decisions* (2nd Ed.). New York: Wiley. https://archive.org/details/judgementchoicep00hogarich

Holmes, Andrew Gary Darwin. (2014). Researcher positionality – a consideration of its influence and place in qualitative research – a new researcher guide. *Shanlax International Journal of Education, 8*(4), 1-10. DOI: https://orcid.org/0000-0002-5147-0761

Hussaini, H. (2020). Role of Qur'anic recitation competition in promoting the study of Qur'anic sciences in Nigeria: Reflections on Bauchi Metropolis. *Interdisciplinary Journal of Education, 3*(1), 1-10. DOI: https://doi.org/10.53449/ije.v3i1.100.

Indra, M.Q. (2019). *Seputar nagham seni baca Al-Qur'an*. Jakarta: Qaf.

Infokini News. (2024, Juni 15). MTQ ke-XXX Provinsi Sulut ternoda, dewan hakim curang, juara 1 jadi juara 2. https://infokini.news/daerah/kotamobagu/mtq-ke-xxx-provinsi-sulut-ternoda-dewan-hakim-curang-juara-1-jadi-juara-2/

Institut Ilmu Al-Qur'an Jakarta (IIQ Jakarta). (2020). Profil Institut Ilmu Al-Qur'an. Retrieved from *https://www.iiq.ac.id*

International Islamic University Malaysia (IIUM). (n.d.). About IRKHS Faculty. Retrieved from *https://www.iium.edu.my*

Jabatan Kemajuan Islam Malaysia. (2019). *Manual Tilawah Al-Quran Antarabangsa*. Kuala Lumpur: Jabatan Kemajuan Islam Malaysia (JAKIM).

Johnson, R. B., and Christensen, L. (2014). *Educational research: Quantitative, qualitative, and mixed approaches*, (5th Ed.). Sage Publications.

Kasim, N. L. A. (2011). Judging behavior and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies, 11*(3), 179-197. ISSN 1675-8021 (In Press). http://www.ukm.my/ppbl/Gema/gemahome.html.

Kementerian Agama Republik Indonesia. (2010). *Al-Qur'an dan tafsirnya*. Jakarta: Lentera Abadi.

Kementerian Agama Republik Indonesia. (2022). *Pedoman penyelenggaraan Musabaqah Tilawatil Quran* (MTQ). Jakarta: Direktorat Jenderal Bimbingan Masyarakat Islam.

Kementerian Agama Republik Indonesia. (n.d.). Kemenag tetapkan 1.998 peserta MTQ Nasional ke-30 di Kalimantan Timur. Diakses pada 6 Januari 2025, dari *https://kemenag.go.id/nasional/kemenag-tetapkan-1-998-peserta-mtq-nasional-ke-30-di-kalimantan-timur-InC78*

Kementerian Agama Republik Indonesia. (2024, Juni 20). Sempurnakan e-MTQ, Menag bertekad bersihkan MTQ dari praktek negatif. https://kemenag.go.id/nasional/sempurnakan-e-mtq-menag-bertekad-bersihkan-mtq-dari-praktek-negatif-kth4px

Kementerian Dalam Negeri. (2025). *Keputusan Menteri Dalam Negeri (Mendagri) Nomor 300.2.2-2138 Tahun 2025.*

Kementerian Agama Republik Indonesia. (2022). *Keputusan Direktur Jenderal Bimbingan Masyarakat Buddha Nomor 78 Tahun 2022 tentang Penetapan Pemenang Mahanitiloka Dhamma Tingkat Nasional Tahun 2022*. Jakarta: Kementerian Agama RI. https://bimasbuddha.kemenag.go.id/mahanitiloka-dhamma-2021-caliadi-mewujudkan-tujuan-pembangunan-nasional-berita-706.html

Kementerian Agama Republik Indonesia. (2018, 28 Oktober). *Menag bersyukur Pesparani pertama tingkat nasional terwujud. Kementerian Agama RI*. https://kemenag.go.id/nasional/menag-bersyukur-pesparani-pertama-tingkat-nasional-terwujud-19dm42

Kementerian Agama Republik Indonesia. (2005). *Peraturan Menteri Agama Nomor 19 Tahun 2005 tentang Pembentukan Lembaga Pengembangan Pesta Paduan Suara Gerejawi Nasional (LPPN).* Jakarta: Kementerian Agama RI.

Kementerian Agama Republik Indonesia. (2016). *Peraturan Menteri Agama Nomor 28 Tahun 2016 tentang Lembaga Pengembangan Dharma Gita*. Jakarta: Kementerian Agama RI.

Kohn, A. (2021). *Punished by rewards: The trouble with gold stars, incentive plans, a's, praise, and other Bribes*. Houghton Mifflin Harcourt.

Kudiya, K., Sumintono, B., Sabana, S., Sachari, A. (2018). Batik artisans' judgment of Batik wax quality and its criteria: An application of the Many-Facets Rasch Model. In Zhang, Q. (Eds) *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings*. Singapore: Springer. DOI: https://doi.org/10.1007/978-981-10-8138-5_3

Landy, F.J., and Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107. DOI: https://doi.org/10.1037/0033-2909.87.1.72.

Levin, K. A. (2006). Study design III: Cross-sectional studies. *Evidence-Based Dentistry, 7*(1), 24–25. DOI: https://doi.org/10.1038/sj.ebd.6400375

Linacre, J. M. (2021). *A user's guide to FACETS: Rasch-Model computer program*. Winsteps.com.

Linacre, J. M. (2022). *Facets computer program for many-facet Rasch measurement*. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2023). *A user's guide to FACETS: Rasch-model computer programs* (Program Manual 3.85.1). Winsteps.com.

Linacre, J. M. (1989). Many-facet Rasch measurement. *Rasch Measurement Transactions, 3*(1), 1-4.

Linacre, J. M. (1989; 1994). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading Massachusetts: Addison-Wesley.

Lunz, M. E., and Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*(4), 425-444. DOI: https://doi.org/10.1177/016327879001300405.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174. DOI: https://doi.org/10.1007/BF02296272.

Masters, G. N. (2016). The partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume one: Models* (pp. 131–144). CRC Press. https://doi.org/10.1201/9781315119144-7

Maryati, M., Prasetyo, Z. K., Wilujeng, I., and Sumintono, B. (2019). Measuring teachers' pedagogical content knowledge using many-facet Rasch model. *Cakrawala Pendidikan, 38*(3), 452–462. DOI: https://doi.org/10.21831/cp.v38i3.26598

Marzano, R. J. (2006). *Classroom assessment and grading that work.* Alexandria, VA: ASCD.

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd Ed.). SAGE Publications.

McKenna, M., Dedrick, R. F., and Goldstein, H. (2022). Development and initial validation of the early elementary writing rubric to inform instruction for kindergarten and first-grade students. *Assessment for Effective Intervention, 47*(4), 220-233. DOI: https://doi.org/10.25384/sage.c.5772932.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed.). New York: Macmillan. https://archive.org/details/educationalmeasu0000unse_3ed/page/n6/mode/1up

Metronews.co. (2024, Juni 18). Temukan kecurangan, sembilan kontingen menolak melanjutkan MTQ ke-44 tingkat provinsi di Balikpapan. https://www.metronews.co/temukan-kecurangan-sembilan-kontingen-menolak-melanjutkan-mtq-ke-44-tingkat-provinsi-di-balikpapan/

Milner IV, H. Richard. (2007). Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher*, *36*(7), 388-400. DOI: https://doi.org/10.3102/0013189X07309471.

Ministry of Islamic Affairs, Kingdom of Saudi Arabia. (2021). *Regulations for the King Abdul Aziz International Quran Competition*. Riyadh: Ministry of Islamic Affairs.

Mitchell, H. F. (2014). *Perception, evaluation and communication of singing voices* (pp. 187–200). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-8851-9_12

Mohamat, R., Sumintono, B., and Abd Hamid, H. S. (2022). Raters' assessment quality in measuring teachers' competency in classroom assessment: Application of many facet rasch model. *Asian Journal of Assessment in Teaching and Learning, 12*(2), 71–88. DOI: https://doi.org/10.37134/ajatel.vol12.2.7.2022

Mohamat, R., Sumintono, B., and Hamid, H. (2022). Analisis kesahan kandungan instrumen kompetensi guru untuk melaksanakan pentaksiran bilik darjah menggunakan model rasch pelbagai faset (Content validity analysis of an instrument to measure teacher's competency for classroom assessment using many facet rasch model). *Jurnal Pendidikan Malaysia, 47*(01). DOI: 10.17576/JPEN-2022-47.01-01

Moskal, B. M., and Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation, 7*(1), 1–7. DOI: https://doi.org/10.7275/q7rm-gg74

M.R, N. M., M. N, M., Chee, J., Mohamad Zin, M. I., Sulaiman, H., and A.G, R. (2015). Penggunaan model pengukuran Rasch many-facet (MFRM) dalam penilaian perkembangan kanak-kanak berasaskan prestasi. *Jurnal Pendidikan Awal Kanak-Kanak Kebangsaan, 4*, 1–16. Retrieved from https://ejournal.upsi.edu.my/index.php/JPAK/article/view/793

Musabaqah.id. (n.d.). MTQN 29 Kalimantan Selatan. Diakses pada 6 Januari 2025, dari *https://musabaqah.id/mtqn-29-kalimantan-selatan/*

Mustaurida, R. (2023, Oktober 18). Tim Balam diduga curang jadi juara umum MTQ ke-50, ini kata Pemkot. IDN Times Lampung. https://lampung.idntimes.com/news/lampung/rohmah-mustaurida/tim-balam-diduga-curang-jadi-juara-umum-mtq-ke-50-ini-kata-pemkot

Myford, C. M., and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422. PubMed ID: 14523257

Nelson, K. (2010). *The Art of Reciting the Qur'an*. Cairo: The American University: Cairo Press.

Nor, M. Z. M., Sumintono, B., Nizam, M. S., and Sani, Z. (2024). Selecting the top artisan woodcraft projects using many facet Rasch measurement (MFRM) model. In Zhang, Q. (Ed.), *Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023)* (pp. 242–250). Atlantis Press. DOI: https://doi.org/10.2991/978-94-6463-494-5_15

Noorhidayati, S., Farihin, H., and Aziz, T. (2021). Melacak sejarah dan penggunaan nagham Arabi di Indonesia. *QOF, 5*(1), 43-58. DOI: https://doi.org/10.30762/Qof.V5i1.

Ogari, E. K. (2020). *Techniques of assessing students' vocal music performance by selected universities in Kenya: Investigating conformity with procedural evaluation frameworks*. Dissertation. https://ir-library.ku.ac.ke/handle/123456789/21496

Ormrod, J. E. (2020). *Human learning* (8th Ed.). Pearson. https://archive.org/details/humanlearning0000ormr_m8r0

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (6th Ed.). Routledge

Plessner, H., and Haar, T. (2006). Sports performance judgments from a social cognitive perspective. *Psychology of Sport and Exercise, 7*(6), 555-575. DOI: https://doi.org/10.1016/j.psychsport.2006.03.007

Popham, W. J. (2019). *Classroom assessment: What teachers need to know* (9th Ed.). Pearson.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. https://archive.org/details/probabilisticmod0000rasc

Rasmussen, A. (2010). *Women, the recited Qur'an, and Islamic music in Indonesia*. University of California Press. https://archive.org/details/womenrecitedqura0000rasm

Ravitch, S. M., and Riggan, M. (2017). *Reason & rigor: How conceptual frameworks guide research* (2nd Ed.). SAGE Publications.

Ryan, R. M., and Deci, E. L. (2020). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. DOI: https://doi.org/10.1007/BF00117714.

Safeí, A. A. (2020). From oral to written da'wah: A study on the development of preaching methods in Indonesia. *International Journal of Psychosocial Rehabilitation, 24*(6), 11473-11481. https://doi.org/10.3390/rel5010179

Safie, S., Yusof, M. Y. Z. bin M., and Ahmad, K. (2021). Credibility and influence of Faridah binti Mat Saman on tarannum institution in Kelantan = Kredibiliti dan pengaruh Faridah binti Mat Saman terhadap institusi tarannum di negeri Kelantan. *AlBayan, 19*(2), 288-309. https://doi.org/10.1163/22321969-12340105.

Salim, M. (2004). *Ilmu nagham Al-Qur'an.* Jakarta: PT Kebayoran Widya Ripta.

Saunders, M., Lewis, P., and Thornhill, A. (2007). *Research methods for business students* (4th ed.). Pearson Education Limited.

Scherbakova A., Engelhard G. Jr., and Bahar A.K. (2025). Development and validation of the scale of aesthetics and creativity in chess. *Front. Psychol*. 16:1545846. DOI: 10.3389/fpsyg.2025.1545846.

Setiawan, M. A., Permatasari, N., and Novitawati. (2024). Meningkatkan motivasi berprestasi peserta Musabaqah Tilawatil Qur'an. *Beujroh: Jurnal Pemberdayaan Dan Pengabdian Pada Masyarakat, 2*(2), 244-257. DOI: https://doi.org/10.61579/Beujroh.V2i2.

Shavelson, R. J., and Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE Publications.

Shihab, M. Q. (2002). *Membumikan Al-Qur'an: Fungsi dan peran wahyu dalam kehidupan masyarakat*. Bandung: Mizan.

Springer, D. G., and Bradley, K. D. (2018). Investigating adjudicator bias in concert band evaluations: An application of the many-facets Rasch model. *Musicae Scientiae, 22*(3), 377–393. DOI: https://doi.org/10.1177/1029864917697782

Strike, Kenneth A. (2006). The ethics of educational research. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in educational research*. (p. 57-73). Routledge.

Styck, K. M., Anthony, C. J., Sandilos, L. E., and DiPerna, J. C. (2020). Examining rater effects on the classroom assessment scoring system. *Child Development, 92*(3), 976-993. DOI: 10.1111/cdev.13460

Sumintono, B., and Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi: Trim Publishing House.

Sumintono, B., and Widhiarso, W. (2013). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial*. Trim Komunika.

Sutter, M., and Kocher, M. G. (2004). Favoritism of agents – The case of referees' home bias. *Journal of Economic Psychology, 25*(4), 461-469. DOI: https://doi.org/10.1016/S0167-4870(03)00013-8

Tennant, A. (2004). Disordered thresholds: An example from the functional independence measure. *Rasch Measurement Transactions, 17*(4), 945–946. https://www.rasch.org/rmt/rmt174.pdf

Tierney, R. D. (2013). Fairness in classroom assessment. In J. H. McMillan (Ed.), *SAGE Handbook of Research on Classroom Assessment* (pp. 125-144). Thousand Oaks, CA: SAGE Publications.

Tim Penyusun. (2023). *Buku pedoman Musabaqah Al-Quran dan Hadits tahun 2023*. Direktorat Penerangan Agama Islam, Direktorat Jenderal Bimbingan Masyarakat Islam, Kementerian Agama Republik Indonesia.

The Royal Islamic Strategic Studies Centre. (2024). The Muslim 500: The world's 500 most influential Muslims, 2024. Amman, Jordan: The Royal Islamic Strategic Studies Centre. Retrieved from https://themuslim500.com/

Totabuan News. (2024, Juni 16). Polemik MTQ ke-XXX Sulut: Dewan hakim dinilai rusak mental peserta. https://totabuan.news/2024/06/polemik-mtq-ke-xxx-sulut-dewan-hakim-dinilai-rusak-mental-peserta/

Totabuan News. (2024, Juni 17). Mengungkap tradisi dugaan kecurangan di MTQ tingkat Provinsi Sulut, dewan hakim selalu jadi dalang. https://www.beritasatu.com/network/totabuannews/207466/mengungkap-tradisi-dugaan-kecurangan-di-mtq-tingkat-provinsi-sulut-dewan-hakim-selalu-jadi-dalang

Traub, R.E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage Publications.

Umm Al-Qura University. (n.d.). Faculty of Da'wah and Usul al-Din. Retrieved from *https://uqu.edu.sa*

Van der Linden, W.J. [Ed.]. (2016). Preface. In *Handbook of item response theory, Vol.2*: Models (ppp. Xvii-xix). Boca Raton, FL: CRC Press.

Wang, J., Ahn, S., and Morgan, S. (2022). Measuring the process of interdisciplinary team collaboration: Creating valid measures using a many-facet Rasch model approach.

Journal of Clinical and Translational Science, 6(1), E134. DOI: 10.1017/cts.2022.472

Wang, J. and Long, H. (2022). Reexamining subjective creativity assessments in science tasks: An application of the rater-mediated assessment framework and many-facet Rasch model. *Psychology of Aesthetics, Creativity, and the Arts*. DOI: https:/doi.org/10/1037/aca0000470.

Wang, P., Coetzee, K., Strachan, A., Monteiro, S., and Cheng, L. (2020). Examining rater performance on the CELBAN speaking: A many-facets Rasch measurement analysis. *Canadian Journal of Applied Linguistics, 23*(2), 73–95. https://doi.org/10.37213/cjal.2020.30436

Wang, Z., and Osterlind, S. J. (2013). Classical test theory. In Teo, T. [Ed.]. *Handbook of quantitative methods for educational research*. Rotterdam: Sense Publishers.

Wesolowski, B. C. (2020). Validity, reliability, and fairness in classroom test. In Parkes, K.A., and Burrack, F. (Ed.), *Developing and applying assessments in the music classroom* (pp. 82-102). New York: Routledge.

Wesolowski, B. C., and Wind, S. A. (2018). Validity, reliability, and fairness in music testing. In T. S. Brophy (Ed.), *The Oxford handbook of assessment, policy, and practice in music education* (pp. 437–460). New York: Oxford University Press.

Wilson, M. (2023). *Constructing measures: An item response modeling approach*, [2nd Ed.]. Taylor & Francis.

Wind, S.A., Engelhard, G., and Wesolowski, B. (2016). Exploring the effects of rater linking designs and rater fit on achievement estimates within the context of music performance assessments. *Educational Assessment, 21*(4), 278–299. https://doi.org/10.1080/10627197.2016.1236676

Wright, B. D., and Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of physical medicine and rehabilitation, 70*(12), 857–860.

Wright, B. D., and Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

Wright, B. D., and Stone, M. H. (2004). *Making measures: Constructing meaning in measurement*. The Phaneron Press, Inc.

Wu, S.M., and Tan, S. (2016). Managing rater effects through the use of FACETS analysis: The case of a university placement test. *Higher Education Research & Development, 35*(2), 380–394. https://doi.org/10.1080/07294360.2015.1087381.

Yahaya, N., Samaila, A., Abdulganiyu, M. L., and Adam, F. S. (2024). Re-Invented tradition: Exploring the contribution of quranic competition on arabic language use and learning in Northern Nigeria. *IJELR: International Journal of Education, Language, and Religion, 6*(2), 131-146. DOI:10.35308/ijelr.v6i2.9495

Yudha, R.P. (2020). Validity and reliability rubric of performance assessment geometry using the Many Facet Rasch Model approach. *EduMa: Mathematics Education Learning and Teaching, 9*(2), 25-34. DOI: 10.24235/eduma.v9i2.7100

Zabidi, Z.M., Sumintono, B., and Zuraidah Abdullah. (2021). Enhancing analytic rigor in qualitative analysis: Developing and testing code scheme using many facet Rasch model. *Quality & Quantity, 56*(3), 1-15. DOI: https://doi.org/10.1007/s11135-021-01152-4

APPENDICES

Appendix 1 Research timeline

Research Stages	Description	Time plan
Preparation of thesis proposals	Compiling chapters I, II, and III	November 2024 – January 2025
Ethical approval	Managing ethical consents for data collection, considering the approval from the faculty	January 2025
Data permission	- Apply for research permission and recommendation to the LPTQ National and LPTQ Banten, for using the instrument and conducting the research - Applying the research permission to the research location - Applying for permits and offering a research assistant, and participants' willingness to participate in this research	January - February 2025
Data collection preparation	- Recruit and explain data collection techniques periodically to research assistants by referring to research protocols - The preparation of data collection began with a workshop for the judges and continued with the preparation of the D-day of the data collection with research assistants	February 2025
Data capture	- Implementation of the Quranic recitation assessment workshop and collection of demographic data or background of the judges	February 2025
	- Collecting the data through Quranic recitation performance at the research location	March 2025

Data cleansing and validation	- Ensuring data is complete, valid, and ready for analysis. - Raw data processing and cleaning.	March 2025
Data analysis	Conducting multi-rater analysis and assessing the quality of Quranic recitation rubric instrument	March – April 2025
Interpretation and discussion of the findings	Compile and write a discussion from the analysis results	April – May 2025
Finalization of research reports	Finalization	May – June 2025

Appendix 2 Request approval for research from the Faculty

Universitas
Islam Internasional
Indonesia

Kementerian Agama Republik Indonesia
Universitas Islam Internasional Indonesia
Jalan Raya Bogor KM. 33.5
Cisalak, Sukmajaya, Depok, Jawa Barat 16416
secretariat@uii.ac.id
www.uii.ac.id

Nomor : 042/Dek.FIP/UIII/UM.02/2/2025 Depok, 12 Februari 2025
Lampiran : -
Perihal : Surat Izin Penggunaan Instrumen dan Permohonan Rekomendasi

Kepada Yth.
Ketua Harian Lembaga Pengembangan Tilawatil Quran (LPTQ) Provinsi Banten,

Assalamu'alaikum Wr. Wb.,

Dengan ini kami menyatakan bahwa mahasiswa di bawah ini:

Nama	: Muhammad Lutfi Assidiqi
NIM	: 04212310002
Fakultas	: Fakultas Ilmu Pendidikan
Program Studi	: Magister Ilmu Pendidikan

sedang melakukan penelitian untuk tesis dengan rincian sebagai berikut:

Judul Tesis	: *Assessing Validity, Reliability, and Fairness of Quranic Recitation Assessment Rubrics Instrument in the Musabaqah Tilawatil Quran (MTQ): Leveraging Many-Facet Rasch Measurement (MFRM)*
Lokasi Penelitian	: [redacted]
Durasi Penelitian	: ± 1 bulan

Sehubungan dengan hal tersebut, kami mengajukan permohonan agar mahasiswa kami dapat diberikan izin untuk menggunakan instrumen penelitian serta memperoleh rekomendasi dari Lembaga Pengembangan Tilawatil Quran (LPTQ) Provinsi Banten demi kelancaran pengumpulan data penelitian.

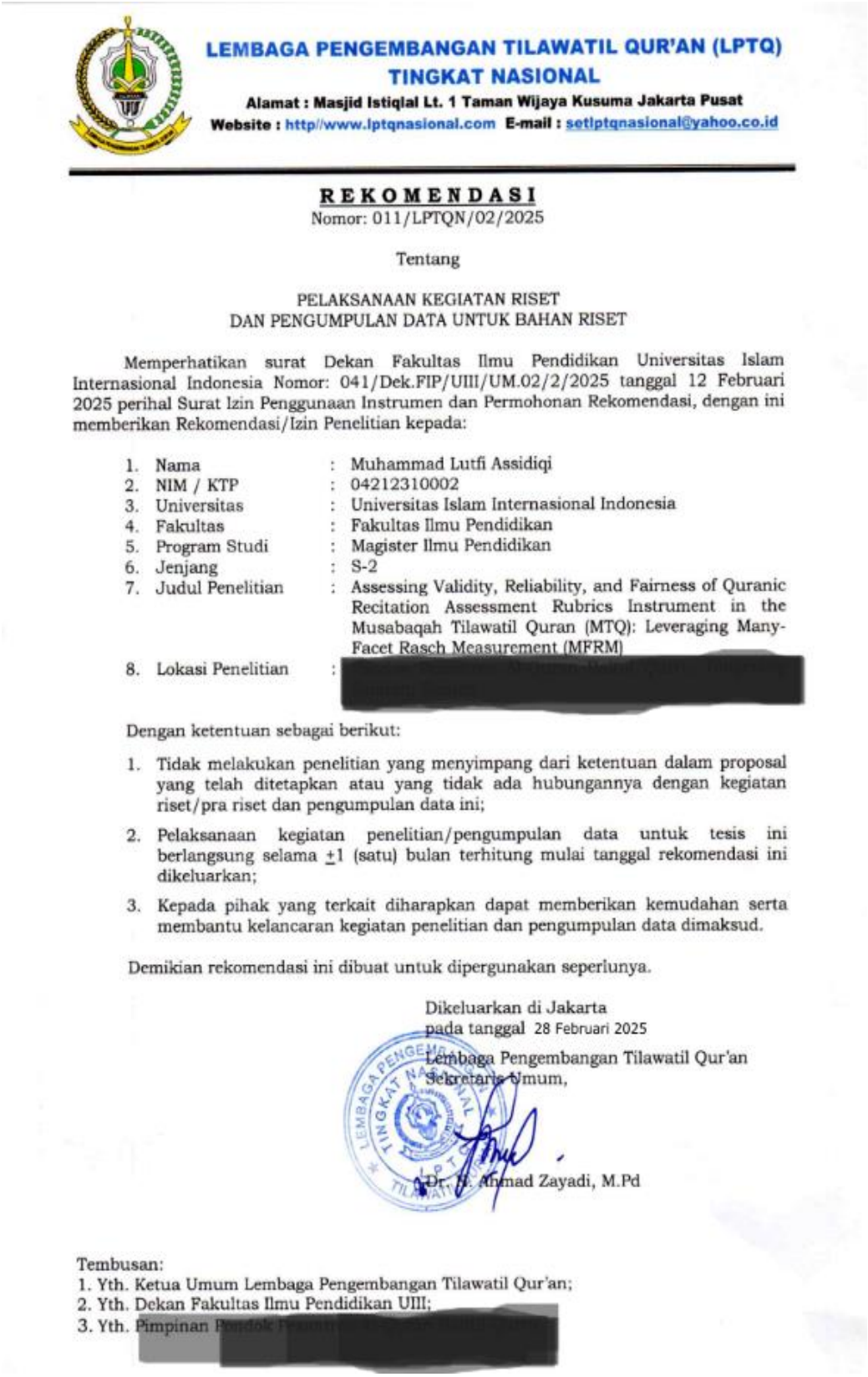
Demikian kami sampaikan, semoga permohonan kami dapat ditindaklanjut sesuai peraturan yang berlaku. Atas perhatian dan kerja sama Ibu/Bapak kami ucapkan terima kasih.

Wassalamu'alaikum Wr. Wb.,

Dengan Hormat,
Dekan Fakultas Ilmu Pendidikan

Prof. Nina Nurmila, PhD

Appendix 3 The letter of research recommendation and Rubric utilized permission from the LPTQ National.



LEMBAGA PENGEMBANGAN TILAWATIL QUR'AN (LPTQ)
TINGKAT NASIONAL
Alamat : Masjid Istiqlal Lt. 1 Taman Wijaya Kusuma Jakarta Pusat
Website : http//www.lptqnasional.com E-mail : setlptqnasional@yahoo.co.id

REKOMENDASI
Nomor: 011/LPTQN/02/2025

Tentang

PELAKSANAAN KEGIATAN RISET
DAN PENGUMPULAN DATA UNTUK BAHAN RISET

Memperhatikan surat Dekan Fakultas Ilmu Pendidikan Universitas Islam Internasional Indonesia Nomor: 041/Dek.FIP/UIII/UM.02/2/2025 tanggal 12 Februari 2025 perihal Surat Izin Penggunaan Instrumen dan Permohonan Rekomendasi, dengan ini memberikan Rekomendasi/Izin Penelitian kepada:

1. Nama : Muhammad Lutfi Assidiqi
2. NIM / KTP : 04212310002
3. Universitas : Universitas Islam Internasional Indonesia
4. Fakultas : Fakultas Ilmu Pendidikan
5. Program Studi : Magister Ilmu Pendidikan
6. Jenjang : S-2
7. Judul Penelitian : Assessing Validity, Reliability, and Fairness of Quranic Recitation Assessment Rubrics Instrument in the Musabaqah Tilawatil Quran (MTQ): Leveraging Many-Facet Rasch Measurement (MFRM)
8. Lokasi Penelitian : [redacted]

Dengan ketentuan sebagai berikut:

1. Tidak melakukan penelitian yang menyimpang dari ketentuan dalam proposal yang telah ditetapkan atau yang tidak ada hubungannya dengan kegiatan riset/pra riset dan pengumpulan data ini;
2. Pelaksanaan kegiatan penelitian/pengumpulan data untuk tesis ini berlangsung selama ±1 (satu) bulan terhitung mulai tanggal rekomendasi ini dikeluarkan;
3. Kepada pihak yang terkait diharapkan dapat memberikan kemudahan serta membantu kelancaran kegiatan penelitian dan pengumpulan data dimaksud.

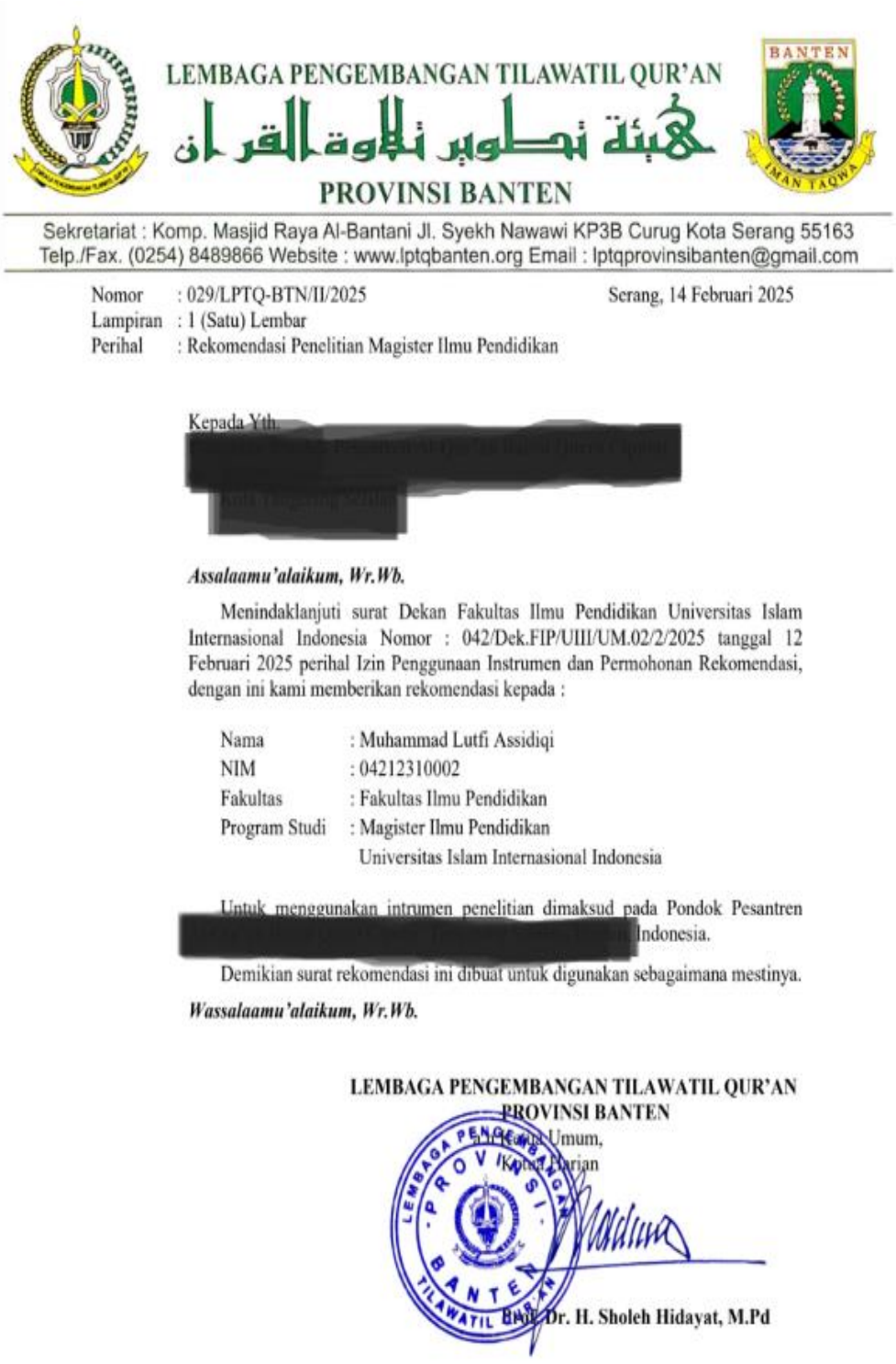
Demikian rekomendasi ini dibuat untuk dipergunakan seperlunya.

Dikeluarkan di Jakarta
pada tanggal 28 Februari 2025
Lembaga Pengembangan Tilawatil Qur'an
Sekretaris Umum,

Dr. H. Ahmad Zayadi, M.Pd

Tembusan:
1. Yth. Ketua Umum Lembaga Pengembangan Tilawatil Qur'an;
2. Yth. Dekan Fakultas Ilmu Pendidikan UIII;
3. Yth. Pimpinan [redacted]

Appendix 4 The letter of research recommendation from the LPTQ Banten Province



LEMBAGA PENGEMBANGAN TILAWATIL QUR'AN

هيئة تطوير تلاوة القرآن

PROVINSI BANTEN

Sekretariat : Komp. Masjid Raya Al-Bantani Jl. Syekh Nawawi KP3B Curug Kota Serang 55163
Telp./Fax. (0254) 8489866 Website : www.lptqbanten.org Email : lptqprovinsibanten@gmail.com

Nomor : 029/LPTQ-BTN/II/2025 Serang, 14 Februari 2025
Lampiran : 1 (Satu) Lembar
Perihal : Rekomendasi Penelitian Magister Ilmu Pendidikan

Kepada Yth.
[redacted]

Assalaamu'alaikum, Wr.Wb.

Menindaklanjuti surat Dekan Fakultas Ilmu Pendidikan Universitas Islam Internasional Indonesia Nomor : 042/Dek.FIP/UIII/UM.02/2/2025 tanggal 12 Februari 2025 perihal Izin Penggunaan Instrumen dan Permohonan Rekomendasi, dengan ini kami memberikan rekomendasi kepada :

Nama : Muhammad Lutfi Assidiqi
NIM : 04212310002
Fakultas : Fakultas Ilmu Pendidikan
Program Studi : Magister Ilmu Pendidikan
Universitas Islam Internasional Indonesia

Untuk menggunakan intrumen penelitian dimaksud pada Pondok Pesantren [redacted] Indonesia.

Demikian surat rekomendasi ini dibuat untuk digunakan sebagaimana mestinya.

Wassalaamu'alaikum, Wr.Wb.

LEMBAGA PENGEMBANGAN TILAWATIL QUR'AN
PROVINSI BANTEN
Ketua Umum,
Ketua Harian

Prof. Dr. H. Sholeh Hidayat, M.Pd

Tembusan :
1. Yth. Ketua Umum LPTQ Provinsi Banten (Sebagai Laporan);
2. Yth. Dekan Fakultas Ilmu Pendidikan Universitas Islam Internasional Indonesia.

Appendix 5 Obtain research permission from the intended institutions

Yayasan Pondok Pesantren Al - Qur'an

[redacted]

Kota Tangerang Selatan . Prov. Banten 15413

SURAT KETERANGAN

Nomor : 020/S.Ket/YPA-BQ/II/2025

Yang bertanda tangan di bawah ini, Ketua Yayasan Pondok Pesantre[redacted]
dengan ini menerangkan bahwa :

Nama Mahasiswa	:	Muhammad Lutfi Assidiqi
NIM	:	04212310002
Fakultas	:	Fakultas Ilmu Pendidikan
Program Studi	:	Magister Ilmu Pendidikan Universitas Islam Internasional Indonesia

Menerima mahasiswa tersebut untuk melaksanakan penelitian di Instansi kami.

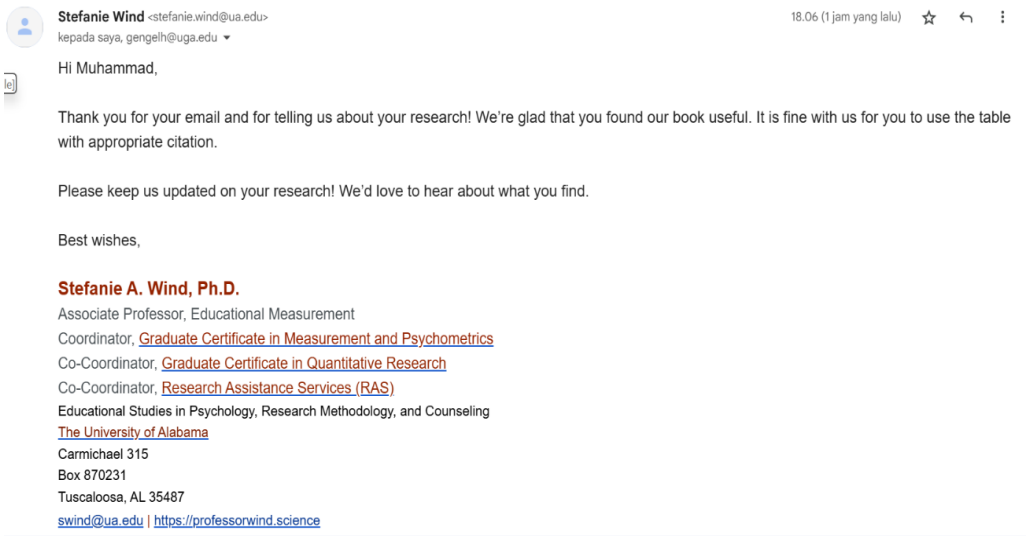
Demikian surat keterangan ini dibuat dengan sebenar-benarnya dan dapat dipergunakan sebagaimana mestinya.

Tangerang Selatan, 21 Februari 2025
Mengetahui,

[redacted]

Dr. Hj. [redacted] M.A
Ketua Yayasan

Appendix 6 Research permission on the adaptation of the table fit category analysis of Mean square error (MSE)



Stefanie Wind <stefanie.wind@ua.edu> 18.06 (1 jam yang lalu)
kepada saya, gengelh@uga.edu

Hi Muhammad,

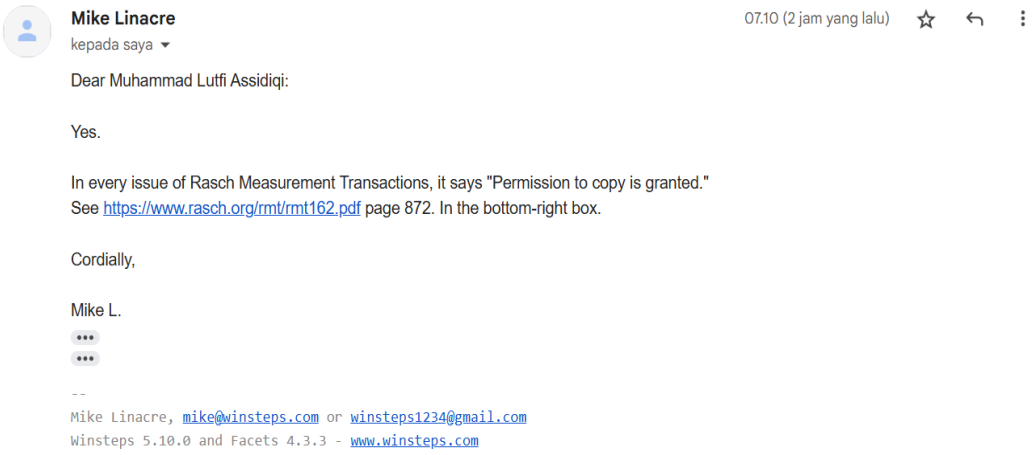
Thank you for your email and for telling us about your research! We're glad that you found our book useful. It is fine with us for you to use the table with appropriate citation.

Please keep us updated on your research! We'd love to hear about what you find.

Best wishes,

Stefanie A. Wind, Ph.D.
Associate Professor, Educational Measurement
Coordinator, Graduate Certificate in Measurement and Psychometrics
Co-Coordinator, Graduate Certificate in Quantitative Research
Co-Coordinator, Research Assistance Services (RAS)
Educational Studies in Psychology, Research Methodology, and Counseling
The University of Alabama
Carmichael 315
Box 870231
Tuscaloosa, AL 35487
swind@ua.edu | https://professorwind.science

Appendix 7 Research permission on the adaptation of the table fit analysis of Z standardized permission



Mike Linacre 07.10 (2 jam yang lalu)
kepada saya

Dear Muhammad Lutfi Assidiqi:

Yes.

In every issue of Rasch Measurement Transactions, it says "Permission to copy is granted."
See https://www.rasch.org/rmt/rmt162.pdf page 872. In the bottom-right box.

Cordially,

Mike L.

--
Mike Linacre, mike@winsteps.com or winsteps1234@gmail.com
Winsteps 5.10.0 and Facets 4.3.3 - www.winsteps.com

Appendix 8 Consent form for research assistants

PERSETUJUAN DAN PERNYATAAN KESEDIAAN

Asisten Peneliti

Yang bertanda tangan di bawah ini:

Nama : [redacted]

Tanggal lahir : 20 - 05 - 2005

Posisi saat ini di Pesantren : Pengurus

Lama tinggal di Pesantren : 5 tahun

Nomor telepon : [redacted]

Dengan ini menyatakan kesediaan untuk berpartisipasi sebagai **Asisten Peneliti** dalam projek penelitian yang berjudul, "Assessing the Validity, Reliability, and Fairness of the Quranic Recitation Assessment Rubric in the MTQ Competition: Leveraging Many-Facet Rasch Measurement" yang dilakukan oleh Muhammad Lutfi Assidiqi.

Saya menyatakan bahwa saya telah membaca, memahami, dan setuju untuk mematuhi semua tanggung jawab, pedoman etika, serta ketentuan kerahasiaan data, yang tercantum dalam dokumen protokol penelitian. Saya juga berkomitmen untuk menjaga integritas proses penelitian dengan menaati batasan dan standar etika yang telah ditetapkan. Saya memahami bahwa pelanggaran terhadap ketentuan ini dapat mengakibatkan pencabutan peran saya dalam penelitian ini serta kemungkinan dilaporkan kepada pihak yang berwenang. Dengan menandatangani dokumen ini, saya menerima semua ketentuan yang telah disebutkan dan menyatakan komitmen saya untuk menjunjung tinggi standar etika tertinggi dalam penelitian ini.

Tangerang Selatan, 18 Februari 2025

Asisten Peneliti

([redacted])

Appendix 9 Research assistant protocol (example pages included)

PROTOKOL PENELITIAN

Untuk thesis berjudul:

Assessing Validity, Reliability, and Fairness of Quranic Recitation Assessment Rubric Instrument in the Musabaqah Tilawatil Quran (MTQ): Leveraging Many-Facet Rasch Measurement (MFRM)

Sebuah Panduan Pengumpulan Data untuk Thesis

Oleh:

Muhammad Lutfi Assidiqi

FACULTY OF EDUCATION

UNIVERSITAS ISLAM INTERNASIONAL INDONESIA

2025

DAFTAR ISI

PANDUAN ASISTEN PENELITI

Guideline ini disusun bagi tim peneliti khususnya research assistant (RA) yang akan turut membantu dalam proses pengambilan data. Berbagai detail komponen dari peran dan tanggung jawab, pelatihan dan panduan etik, pengelolaan data, batasan dan larangan, serta kode etik dan kepatuhan. Berikut ini adalah penjelasan secara rinci:

1. Peran dan Tanggung Jawab Asisten Peneliti (RA)

Asisten Peneliti (RA) bertanggung jawab dalam mendukung proses pengumpulan data di bawah supervisi peneliti utama. Tugas RA meliputi:

- Membantu peneliti utama dalam mendistribusikan, menjelaskan detail peran setiap partisipan, dan mengumpulkan lembar persetujuan dan kesediaan sebagai partisipan penelitian (*terlampir*).
- Membantu peneliti utama dalam mengorganisir seluruh rangkaian kegiatan pengumpulan data (*terlampir*).
- Mendistibusikan rubrik penilaian tilawah Al-Qur'an (Quranic recitation) kepada juri para juri.
- Membagikan Maqro' (Tugas bacaan) kepada para penampil (siswa) dua hari sebelum dilaksanakannya kegiatan utama (*terlampir*)
- Mengumpulkan dan mengorganisir data penilaian dari para juri.
- Memastikan semua lembar penilaian terisi dengan lengkap dan benar.
- Berkomunikasi dengan peserta (juri) untuk mengklarifikasi pertanyaan terkait proses penilaian.
- Berkomunikasi dengan peserta (siswa) dalam pelaksanaan penampilan tilawah Al-Quran (Quranic recitation)
- Melaporkan inkonsistensi data, data yang hilang, atau masalah etis kepada peneliti utama.
- Menyerahkan data yang telah dikumpulkan sesuai dengan jadwal penelitian.

2. Pelatihan dan Panduan Etika untuk Asisten Peneliti

Sebelum pengumpulan data, RA wajib mengikuti sesi pelatihan dari peneliti utama yang mencakup:

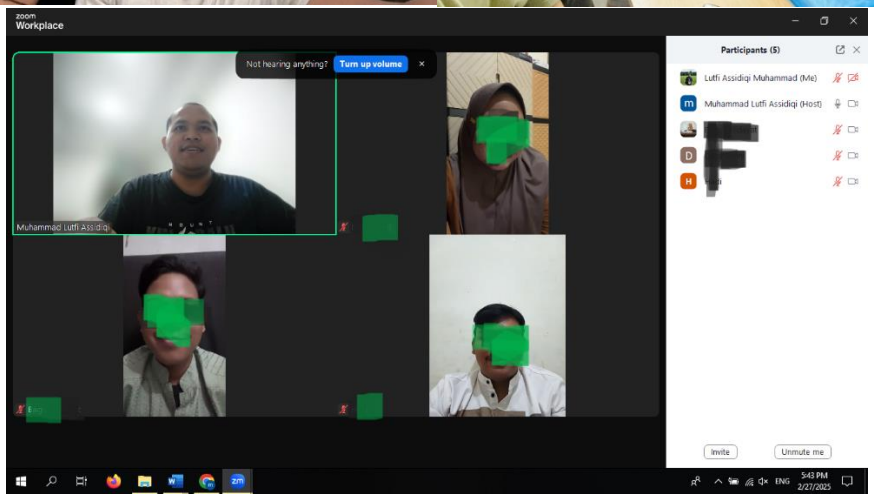
- Etika penelitian dan kerahasiaan (menjaga keamanan dan privasi data juri).
- Prosedur standar pengumpulan data (untuk memastikan keakuratan dan menghindari bias).
- Komunikasi yang tepat dengan peserta (profesional, netral, dan tidak mempengaruhi hasil).
- Keamanan data (cara menyimpan dan mentransfer data dengan aman).
- Semua RA diwajibkan menandatangani Perjanjian Kerahasiaan sebelum mulai bertugas.

3. Pengelolaan Data oleh Asisten Peneliti

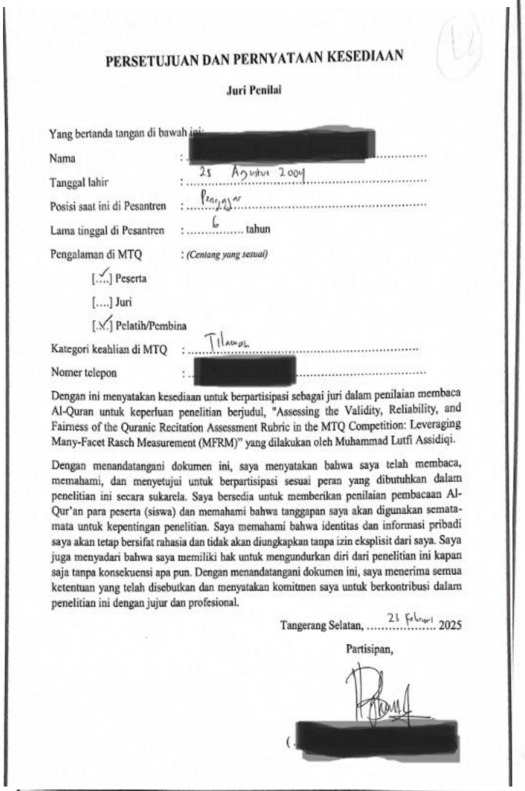
MAQRO' UNTUK PEMBACAAN AL-QURAN PARA PESERTA

No	Nama Surah	Ayat	No. Surah	Hal	Ayat
1	Al-Baqarah	249	2	41	فَلَمَّا فَصَلَ طَالُوتُ بِالْجُنُودِ
2	Al-Baqarah	258	2	43	أَلَمْ تَرَ إِلَى الَّذِي حَاجَّ إِبْرَاهِيمَ فِي رَبِّهِ
3	Al-Baqarah	261	2	44	مَثَلُ الَّذِينَ يُنْفِقُونَ أَمْوَالَهُمْ فِي سَبِيلِ اللَّهِ
4	Al-Baqarah	274	2	46	الَّذِينَ يُنْفِقُونَ أَمْوَالَهُمْ بِاللَّيْلِ وَالنَّهَارِ
5	Ali Imran	110	3	64	كُنْتُمْ خَيْرَ أُمَّةٍ أُخْرِجَتْ لِلنَّاسِ
6	Ali Imran	121	3	65	وَإِذْ غَدَوْتَ مِنْ أَهْلِكَ تُبَوِّئُ الْمُؤْمِنِينَ
7	Ali Imran	149	3	69	يَا أَيُّهَا الَّذِينَ آمَنُوا إِنْ تُطِيعُوا الَّذِينَ كَفَرُوا
8	Ali Imran	156	3	70	يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَكُونُوا كَالَّذِينَ كَفَرُوا

Appendix 10 Research assistants training 1, 2, and 3



Appendix 11 Sample of Judges' participant consent in each group's dimension



PERSETUJUAN DAN PERNYATAAN KESEDIAAN

Juri Penilai

Yang bertanda tangan di bawah [redacted]

Nama : [redacted]

Tanggal lahir : 25 Agustus 2004

Posisi saat ini di Pesantren : Pengajar

Lama tinggal di Pesantren : 6 tahun

Pengalaman di MTQ : *(Centang yang sesuai)*

[✓] Peserta

[....] Juri

[✓] Pelatih/Pembina

Kategori keahlian di MTQ : Tilawah

Nomer telepon : [redacted]

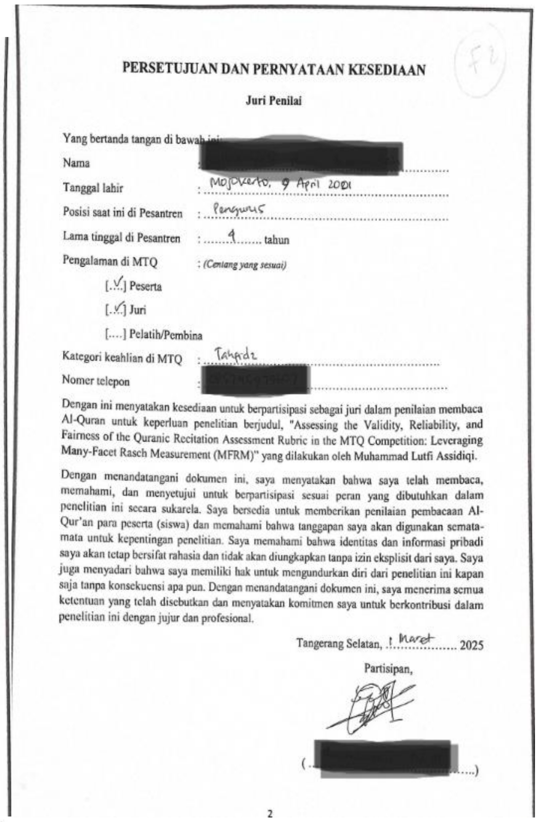
Dengan ini menyatakan kesediaan untuk berpartisipasi sebagai juri dalam penilaian membaca Al-Quran untuk keperluan penelitian berjudul, "Assessing the Validity, Reliability, and Fairness of the Quranic Recitation Assessment Rubric in the MTQ Competition: Leveraging Many-Facet Rasch Measurement (MFRM)" yang dilakukan oleh Muhammad Lutfi Assidiqi.

Dengan menandatangani dokumen ini, saya menyatakan bahwa saya telah membaca, memahami, dan menyetujui untuk berpartisipasi sesuai peran yang dibutuhkan dalam penelitian ini secara sukarela. Saya bersedia untuk memberikan penilaian pembacaan Al-Qur'an para peserta (siswa) dan memahami bahwa tanggapan saya akan digunakan semata-mata untuk kepentingan penelitian. Saya memahami bahwa identitas dan informasi pribadi saya akan tetap bersifat rahasia dan tidak akan diungkapkan tanpa izin eksplisit dari saya. Saya juga menyadari bahwa saya memiliki hak untuk mengundurkan diri dari penelitian ini kapan saja tanpa konsekuensi apa pun. Dengan menandatangani dokumen ini, saya menerima semua ketentuan yang telah disebutkan dan menyatakan komitmen saya untuk berkontribusi dalam penelitian ini dengan jujur dan profesional.

Tangerang Selatan, 21 Februari 2025

Partisipan,

([redacted])



PERSETUJUAN DAN PERNYATAAN KESEDIAAN

Juri Penilai

Yang bertanda tangan di bawah [redacted]

Nama : [redacted]

Tanggal lahir : Mojokerto, 9 April 2001

Posisi saat ini di Pesantren : Pengurus

Lama tinggal di Pesantren : 4 tahun

Pengalaman di MTQ : *(Centang yang sesuai)*

[✓] Peserta

[✓] Juri

[....] Pelatih/Pembina

Kategori keahlian di MTQ : Tahfidz

Nomer telepon : [redacted]

Dengan ini menyatakan kesediaan untuk berpartisipasi sebagai juri dalam penilaian membaca Al-Quran untuk keperluan penelitian berjudul, "Assessing the Validity, Reliability, and Fairness of the Quranic Recitation Assessment Rubric in the MTQ Competition: Leveraging Many-Facet Rasch Measurement (MFRM)" yang dilakukan oleh Muhammad Lutfi Assidiqi.

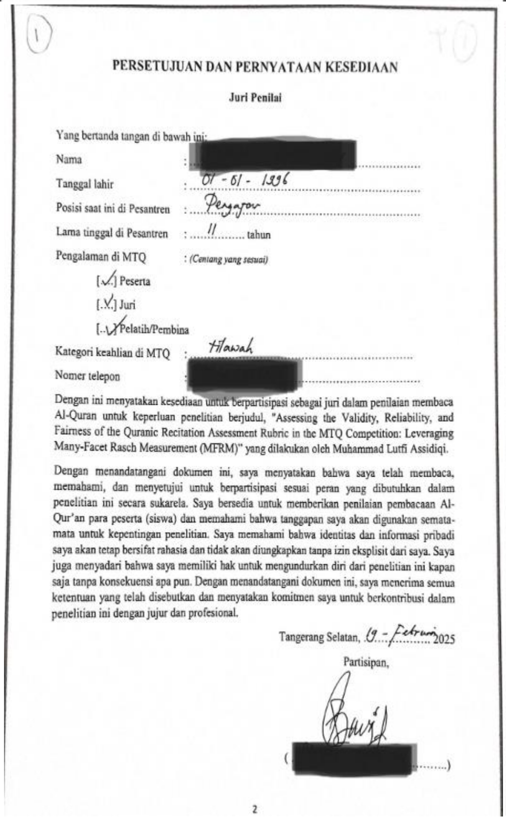
Dengan menandatangani dokumen ini, saya menyatakan bahwa saya telah membaca, memahami, dan menyetujui untuk berpartisipasi sesuai peran yang dibutuhkan dalam penelitian ini secara sukarela. Saya bersedia untuk memberikan penilaian pembacaan Al-Qur'an para peserta (siswa) dan memahami bahwa tanggapan saya akan digunakan semata-mata untuk kepentingan penelitian. Saya memahami bahwa identitas dan informasi pribadi saya akan tetap bersifat rahasia dan tidak akan diungkapkan tanpa izin eksplisit dari saya. Saya juga menyadari bahwa saya memiliki hak untuk mengundurkan diri dari penelitian ini kapan saja tanpa konsekuensi apa pun. Dengan menandatangani dokumen ini, saya menerima semua ketentuan yang telah disebutkan dan menyatakan komitmen saya untuk berkontribusi dalam penelitian ini dengan jujur dan profesional.

Tangerang Selatan, 1 Maret 2025

Partisipan,

([redacted])

2



PERSETUJUAN DAN PERNYATAAN KESEDIAAN

Juri Penilai

Yang bertanda tangan di bawah [redacted]

Nama : [redacted]

Tanggal lahir : 07 - 01 - 1996

Posisi saat ini di Pesantren : Pengajar

Lama tinggal di Pesantren : 11 tahun

Pengalaman di MTQ : *(Centang yang sesuai)*

[✓] Peserta

[✓] Juri

[✓] Pelatih/Pembina

Kategori keahlian di MTQ : Tilawah

Nomer telepon : [redacted]

Dengan ini menyatakan kesediaan untuk berpartisipasi sebagai juri dalam penilaian membaca Al-Quran untuk keperluan penelitian berjudul, "Assessing the Validity, Reliability, and Fairness of the Quranic Recitation Assessment Rubric in the MTQ Competition: Leveraging Many-Facet Rasch Measurement (MFRM)" yang dilakukan oleh Muhammad Lutfi Assidiqi.

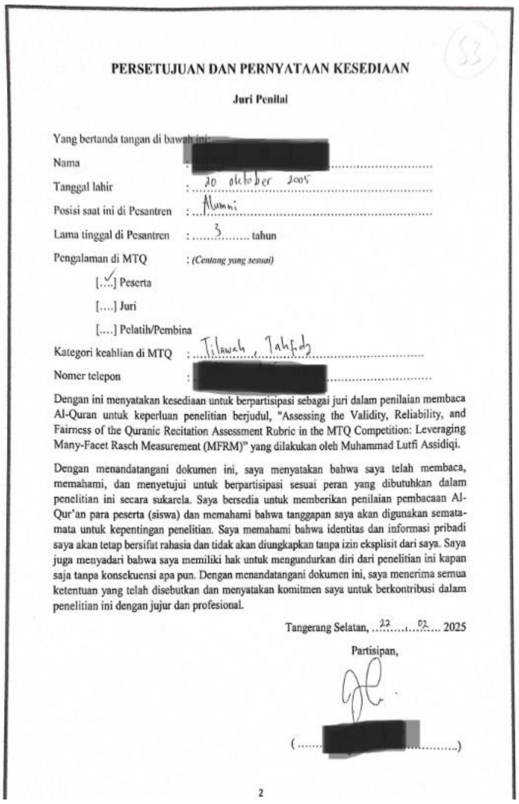
Dengan menandatangani dokumen ini, saya menyatakan bahwa saya telah membaca, memahami, dan menyetujui untuk berpartisipasi sesuai peran yang dibutuhkan dalam penelitian ini secara sukarela. Saya bersedia untuk memberikan penilaian pembacaan Al-Qur'an para peserta (siswa) dan memahami bahwa tanggapan saya akan digunakan semata-mata untuk kepentingan penelitian. Saya memahami bahwa identitas dan informasi pribadi saya akan tetap bersifat rahasia dan tidak akan diungkapkan tanpa izin eksplisit dari saya. Saya juga menyadari bahwa saya memiliki hak untuk mengundurkan diri dari penelitian ini kapan saja tanpa konsekuensi apa pun. Dengan menandatangani dokumen ini, saya menerima semua ketentuan yang telah disebutkan dan menyatakan komitmen saya untuk berkontribusi dalam penelitian ini dengan jujur dan profesional.

Tangerang Selatan, 19 - Februari 2025

Partisipan,

([redacted])

2



PERSETUJUAN DAN PERNYATAAN KESEDIAAN

Juri Penilai

Yang bertanda tangan di bawah [redacted]

Nama : [redacted]

Tanggal lahir : 20 oktober 2005

Posisi saat ini di Pesantren : Alumni

Lama tinggal di Pesantren : 3 tahun

Pengalaman di MTQ : *(Centang yang sesuai)*

[✓] Peserta

[....] Juri

[....] Pelatih/Pembina

Kategori keahlian di MTQ : Tilawah, Tahfidz

Nomer telepon : [redacted]

Dengan ini menyatakan kesediaan untuk berpartisipasi sebagai juri dalam penilaian membaca Al-Quran untuk keperluan penelitian berjudul, "Assessing the Validity, Reliability, and Fairness of the Quranic Recitation Assessment Rubric in the MTQ Competition: Leveraging Many-Facet Rasch Measurement (MFRM)" yang dilakukan oleh Muhammad Lutfi Assidiqi.

Dengan menandatangani dokumen ini, saya menyatakan bahwa saya telah membaca, memahami, dan menyetujui untuk berpartisipasi sesuai peran yang dibutuhkan dalam penelitian ini secara sukarela. Saya bersedia untuk memberikan penilaian pembacaan Al-Qur'an para peserta (siswa) dan memahami bahwa tanggapan saya akan digunakan semata-mata untuk kepentingan penelitian. Saya memahami bahwa identitas dan informasi pribadi saya akan tetap bersifat rahasia dan tidak akan diungkapkan tanpa izin eksplisit dari saya. Saya juga menyadari bahwa saya memiliki hak untuk mengundurkan diri dari penelitian ini kapan saja tanpa konsekuensi apa pun. Dengan menandatangani dokumen ini, saya menerima semua ketentuan yang telah disebutkan dan menyatakan komitmen saya untuk berkontribusi dalam penelitian ini dengan jujur dan profesional.

Tangerang Selatan, 22 . 02 . 2025

Partisipan,

([redacted])

2

Appendix 12 Sample of student's performance consent

PERSETUJUAN DAN PERNYATAAN KESEDIAAN

Siswa sebagai Penampil

Saya yang bertanda tangan di bawah ini menyatakan kesediaan saya untuk berpartisipasi sebagai peserta tilawah Al-Quran (Quranic recitation) dalam penelitian berjudul, "Assessing the Validity, Reliability, and Fairness of the Quranic Recitation Assessment Rubric in the MTQ Competition: Leveraging Many-Facet Rasch Measurement (MFRM)" yang dilakukan oleh Muhammad Lutfi Assidiqi.

Detail tugas partisipan (Siswa sebagai penampil):

Sebagai bagian dari penelitian ini, peserta akan diminta untuk melantunkan tilawah Al-Qur'an sesuai dengan maqra' (tugas yang dibaca) yang telah ditentukan dan diberikan sebelumnya (*terlampir*). Pelaksanaan tilawah akan dilakukan dalam suasana yang menyerupai standar pelaksanaan kompetisi MTQ. Setiap peserta akan dinilai oleh panel juri menggunakan rubrik penilaian terstruktur yang mencakup aspek penilaian, *tajwid, fasahah*, suara, dan lagu. Seluruh proses akan diawasi, didokumentasi, dan diobservasi untuk keperluan penelitian.

Kerahasiaan dan Etika Penelitian:

- Semua data yang dikumpulkan akan dijaga kerahasiaannya secara ketat.
- Identitas dan catatan kinerja peserta akan tetap anonim dalam publikasi atau laporan penelitian.
- Informasi yang diperoleh hanya akan digunakan untuk keperluan penelitian akademik dan evaluasi.
- Partisipasi dalam penelitian ini bersifat sukarela, dan peserta berhak mengundurkan diri kapan saja tanpa konsekuensi apa pun.

Dengan menandatangani dokumen ini, saya menyatakan bahwa saya telah membaca, memahami, dan secara sukarela setuju untuk berpartisipasi dalam penelitian ini. Saya memahami bahwa hasil penilaian saya hanya akan digunakan untuk tujuan penelitian dan akan tetap bersifat rahasia. Saya juga menyadari bahwa saya memiliki hak untuk mengundurkan diri kapan saja tanpa konsekuensi apa pun. Saya juga memahami bahwa partisipasi saya dalam penelitian ini tidak akan memengaruhi status maupun hasil akademik saya. Dengan menandatangani di bawah ini, saya mengonfirmasi partisipasi saya secara sukarela dan menyetujui semua ketentuan yang disebutkan di atas.

No	Nama	Kelas	Umur	Keterangan (Bersedia/ tidak)	Signature
1.	[redacted]	IX	15	Bersedia	
2.	[redacted]	IX	14	Bersedia	
3.	[redacted]	IX	14	Bersedia	
4.	[redacted]	VII	12	Bersedia	
5.	[redacted]	XII	19	Bersedia	

3

… SEBAGAI PESERTA PENILAIAN TILAWAH AL-QURAN (QURANIC RECITATION) DALAM KEGIATAN PENELITIAN

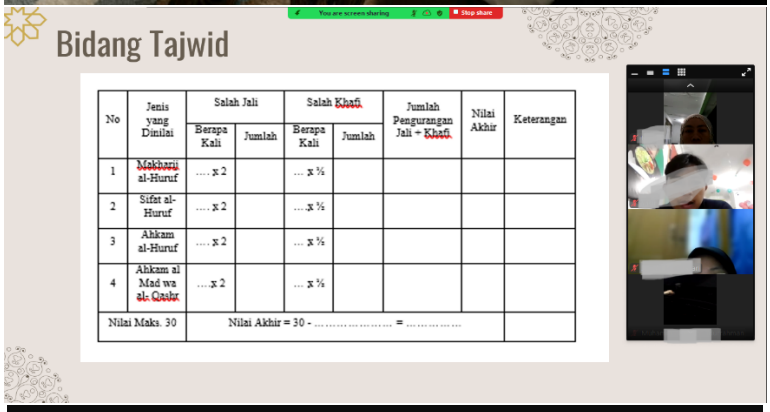
	Asal Daerah	Umur	Kelas	Apakah Anda pernah mengikuti MTQ cabang Tilawah Al-Quran? Jika pernah, berapa kali Anda mengikuti MTQ? 1-5x / 6-10x / > 10x *Contoh: Ya, 6-10 kali atau Tidak, belum pernah*
…nuny	Jakarta	15	9	1 - 5x
…yal.	Bogor	17	11	Belum pernah.
	jakarta	13	7	Belum pernah
	Lombok	17	11	1 - 5x
	Palangkaraya	16	10	6 - 10x
	Bengkulu	17	12	>10x
…afib	Gorontalo	16	11	1 - 5x
	Manado	14	9	6 - 10x
	Tangerang Selatan	15	10	1 - 5x
	Jakarta	15	9	1 - 5x
	Gorontalo	16	11	>10x

35.	[redacted]	XI	16	Bersedia	
36.	[redacted]	X	16	Bersedia	
37.	[redacted]	X	15	Bersedia	
38.	[redacted]	XII	17	Bersedia	
39.	[redacted]	XII	17	Bersedia	
40.	[redacted]	XII	17	Bersedia	
41	[redacted]	X	15	Bersedia	
42	[redacted]	VIII	14	Bersedia	
43	[redacted]	VIII	14	Bersedia	
44	[redacted]	XI	17	Bersedia	
45	[redacted]	IX	14	Bersedia	
46	[redacted]	VII	13	Bersedia	
47	[redacted]	VII	12	Bersedia	
48.	[redacted]	IX	14	Bersedia	
49.	[redacted]	XI	16	Bersedia	
50.	[redacted]	X	16	Bersedia	

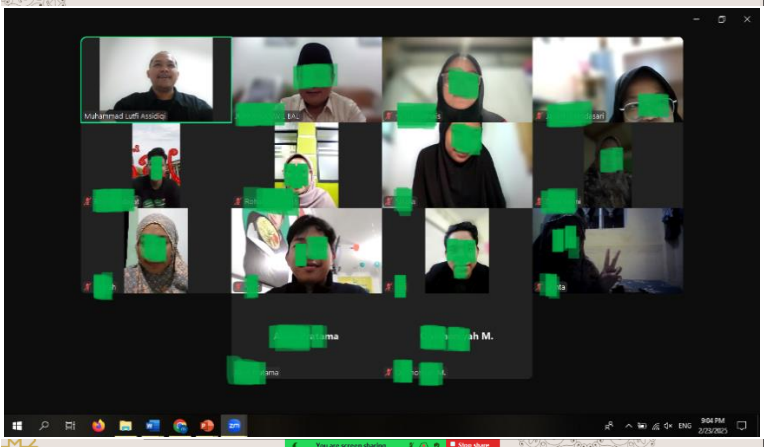
6.	[redacted]	XII	18	Bersedia	
7.	[redacted]	X	15	Bersedia	
8.	[redacted]	X	15	Bersedia	
9.	[redacted]	X	15	Bersedia	
10.	[redacted]	XI	16	Bersedia	
11	[redacted]	XI	16	bersedia	
12.	[redacted]	X	15	Bersedia	
13.	[redacted]	VII	13	bersedia	
14.	[redacted]	VII	13	Bersedia	
15.	[redacted]	VII	13	Bersedia	
16	[redacted]	IX	14	Bersedia	
17.	[redacted]	VII	13	Bersedia	
18	[redacted]	VII	13	bersedia	
19	[redacted]	XII	12	bersedia	
20.	[redacted]	VIII	13	Bersedia	
21	[redacted]	X	15	Bersedia	
22.	[redacted]	IX	14	Bersedia	
23.	[redacted]	X	16	Bersedia	
24.	[redacted]	X	15	Bersedia	
25.	[redacted]	VIII	13	Bersedia	
26	[redacted]	X	15	bersedia	
27.	[redacted]	X	16	Bersedia	
28	[redacted]	XI	17	Bersedia	
29	[redacted]	VIII	14	Bersedia	
30	[redacted]	XII	17	Bersedia	
31.	[redacted]	X	17	Bersedia	
32.	[redacted]	XI	17	Bersedia	
33	[redacted]	XII	18	Bersedia	
34	[redacted]	X	16	Bersedia	

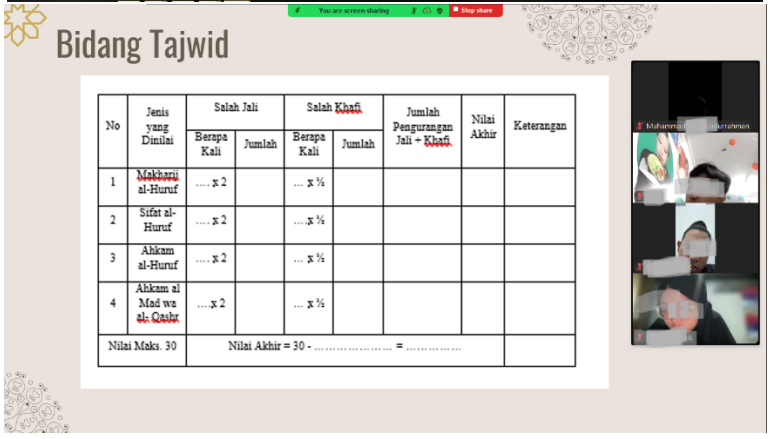
4

Appendix 13 Workshop on filling out the Quranic recitation assessment rubric



No	Jenis yang Dinilai	Salah Jali		Salah Khafi		Jumlah Pengurangan Jali + Khafi	Nilai Akhir	Keterangan
		Berapa Kali	Jumlah	Berapa Kali	Jumlah			
1	Makharij al-Huruf	.... x 2		... x ½				
2	Sifat al-Huruf	.... x 2		... x ½				
3	Ahkam al-Huruf	.... x 2		... x ½				
4	Ahkam al Mad wa al- Qashr	....x 2		... x ½				
Nilai Maks. 30		Nilai Akhir = 30 - ..... = ...						

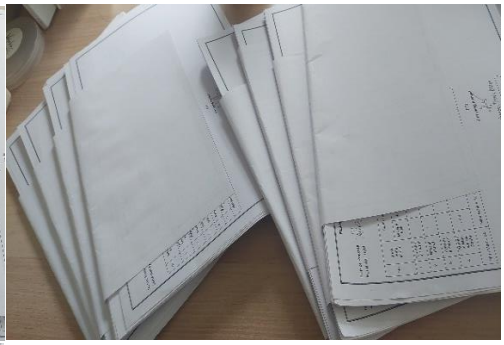


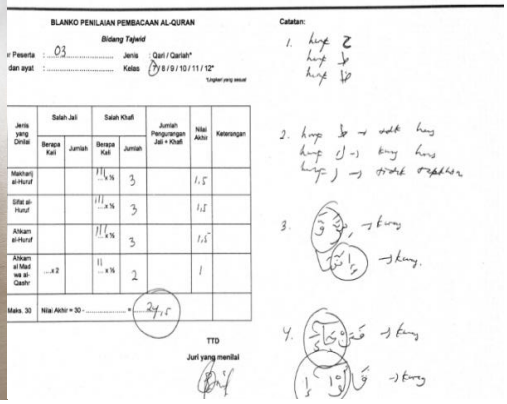


No	Jenis yang Dinilai	Salah Jali		Salah Khafi		Jumlah Pengurangan Jali + Khafi	Nilai Akhir	Keterangan
		Berapa Kali	Jumlah	Berapa Kali	Jumlah			
1	Makharij al-Huruf	.... x 2		... x ½				
2	Sifat al-Huruf	.... x 2		... x ½				
3	Ahkam al-Huruf	.... x 2		... x ½				
4	Ahkam al Mad wa al- Qashr	....x 2		... x ½				
Nilai Maks. 30		Nilai Akhir = 30 - ..... = ...						

Appendix 14 Data collection documentation

Appendix 15 The sample of data from the assessment rubric





BLANKO PENILAIAN PEMBACAAN AL-QURAN

Bidang Tajwid

Nomor Peserta : 03 Jenis : Qari / Qariah*

Surat dan ayat : ..... Kelas : 7 / 8 / 9 / 10 / 11 / 12*

*Lingkari yang sesuai

Jenis yang Dinilai	Salah Jali		Salah Khafi		Jumlah Pengurangan Jali + Khafi	Nilai Akhir	Keterangan
	Berapa Kali	Jumlah	Berapa Kali	Jumlah			
Makharij al-Huruf			... x ½	3		1,5	
Sifat al-Huruf			... x ½	3		1,5	
Ahkam al-Huruf			... x ½	3		1,5	
Ahkam al Mad wa al-Qashr	...x 2		... x ½	2		1	
Maks. 30	Nilai Akhir = 30 - ..... =					24,5	

TTD

Juri yang menilai

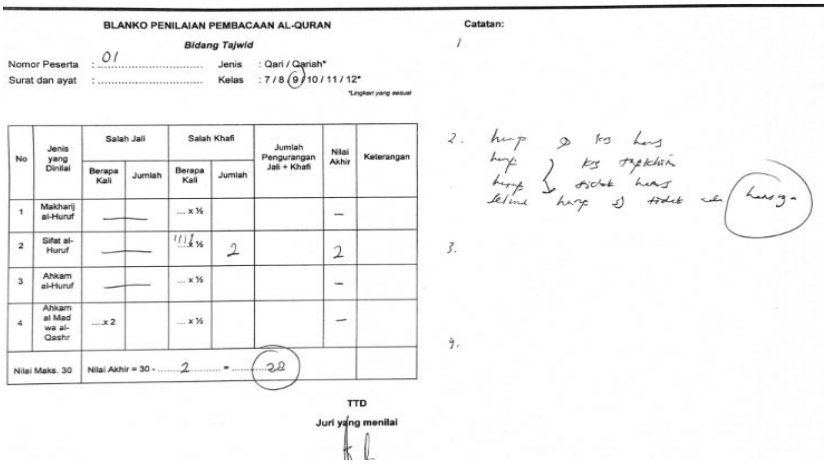
Catatan:

1. huruf ... kurang
huruf ... kurang
huruf ... kurang

2. huruf ... tidak hilang
huruf ... kurang hilang
huruf ... tidak terpakai

3. ... → kurang
... → kurang.

4. ... → kurang
... → kurang



BLANKO PENILAIAN PEMBACAAN AL-QURAN

Bidang Tajwid

Nomor Peserta : 01 Jenis : Qari / Qariah*

Surat dan ayat : ..... Kelas : 7 / 8 / 9 / 10 / 11 / 12*

*Lingkari yang sesuai

No	Jenis yang Dinilai	Salah Jali		Salah Khafi		Jumlah Pengurangan Jali + Khafi	Nilai Akhir	Keterangan
		Berapa Kali	Jumlah	Berapa Kali	Jumlah			
1	Makharij al-Huruf	—		... x ½			–	
2	Sifat al-Huruf	—		IIII x ½	2		2	
3	Ahkam al-Huruf	—		... x ½			–	
4	Ahkam al Mad wa al-Qashr	...x 2		... x ½			–	
Nilai Maks. 30	Nilai Akhir = 30 - ..... 2 ..... =						22	

TTD

Juri yang menilai

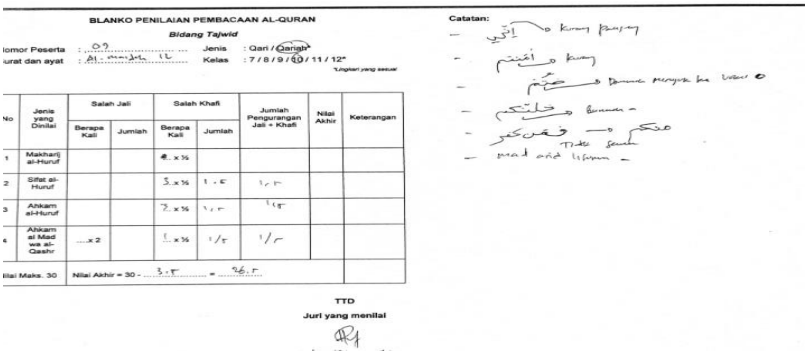
Catatan:

1.

2. huruf ... kurang
huruf ... kurang terpakai
huruf ... tidak hilang
selain huruf ... tidak ... hilang

3.

4.



BLANKO PENILAIAN PEMBACAAN AL-QURAN

Bidang Tajwid

Nomor Peserta : 09 Jenis : Qari / Qariah*

Surat dan ayat : Al-Insyirah 1-6 Kelas : 7 / 8 / 9 / 10 / 11 / 12*

*Lingkari yang sesuai

No	Jenis yang Dinilai	Salah Jali		Salah Khafi		Jumlah Pengurangan Jali + Khafi	Nilai Akhir	Keterangan
		Berapa Kali	Jumlah	Berapa Kali	Jumlah			
1	Makharij al-Huruf			4 x ½				
2	Sifat al-Huruf			3 x ½	1,5			
3	Ahkam al-Huruf			3 x ½	1,5			
4	Ahkam al Mad wa al-Qashr	...x 2		1 x ½	1/2			
Nilai Maks. 30	Nilai Akhir = 30 - 3,5 =						26,5	

TTD

Juri yang menilai

Catatan:

- ... kurang panjang
- ... kurang
- ... Dammah menyatu ke huruf ...
- ... Berhenti
- ... Tidak sesuai
- mad ada lisirnya